知识规则化描述研究

咨询报告

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | |
|  | | 编 写 |  |
| 校 对 |  |
| 审 核 |  |
| 批 准 |  |
|  | | | | |

目录

[**1前言** 1](#_Toc84500719)

[**2知识分类及描述方法选用规范** 3](#_Toc84500720)

[2.1知识分类规范 3](#_Toc84500721)

[2.2知识规则化描述方法及选用规范 3](#_Toc84500722)

[2.2.1产生式规则 3](#_Toc84500723)

[2.2.2知识图谱 5](#_Toc84500724)

[2.2.3框架法 6](#_Toc84500725)

[2.2.4谓词逻辑 8](#_Toc84500726)

[2.2.5基于人工智能的方法 8](#_Toc84500727)

[2.2.6知识规则化描述方法选用规范 9](#_Toc84500728)

[2.3知识分类示例 9](#_Toc84500729)

[**3专家知识抽取与规则化描述** 14](#_Toc84500730)

[3.1知识抽取方式 14](#_Toc84500731)

[3.1.1人工抽取 14](#_Toc84500732)

[3.1.2半自动抽取 16](#_Toc84500733)

[3.1.3自动抽取 18](#_Toc84500734)

[3.2知识规则化 20](#_Toc84500735)

[3.2.1产生式规则 20](#_Toc84500736)

[3.2.2知识图谱 21](#_Toc84500737)

[3.2.3框架法 24](#_Toc84500738)

[3.2.4谓词逻辑 25](#_Toc84500739)

[3.2.5基于人工智能的描述 27](#_Toc84500740)

[3.3存储规范 29](#_Toc84500741)

[3.3.1 图知识的存储 29](#_Toc84500742)

[3.3.2 表知识的存储 31](#_Toc84500743)

[3.3.3 存储工具推荐 33](#_Toc84500744)

[3.4知识规则化示例 33](#_Toc84500745)

[**4知识应用规范** 36](#_Toc84500746)

[4.1知识库设计 36](#_Toc84500747)

[4.1.1源知识数据管理 36](#_Toc84500748)

[4.1.2规则名和事实名的编码规范 37](#_Toc84500749)

[4.2知识应用评价准则 38](#_Toc84500750)

[4.3知识维护流程 39](#_Toc84500751)

[4.3.1一致性检测 39](#_Toc84500752)

[4.3.2完备性维护 43](#_Toc84500753)

[4.3.3知识更新 43](#_Toc84500754)

[4.3.4知识备份 46](#_Toc84500755)

[4.4知识维护示例 48](#_Toc84500756)

[**5初步应用建议** 54](#_Toc84500757)

[5.1现有知识类型分析 54](#_Toc84500758)

[5.2知识抽取方式分析 54](#_Toc84500759)

[5.3规则化方法选型 55](#_Toc84500760)

[5.3.1故障树知识 55](#_Toc84500761)

[5.3.2预案知识 56](#_Toc84500762)

[5.4知识存储 57](#_Toc84500763)

[5.4.1故障树知识数据存储形式 57](#_Toc84500764)

[5.4.2预案知识数据存储形式 57](#_Toc84500765)

[5.5知识维护 58](#_Toc84500766)

[5.5.1规则知识的维护 59](#_Toc84500767)

[5.5.2知识图谱的维护 59](#_Toc84500768)

[5.5.3预案知识的维护 59](#_Toc84500769)

[**参考文献** 60](#_Toc84500770)

[**附件1风洞设备知识管理案例** 61](#_Toc84500771)

[F1.1风洞设备故障诊断知识管理案例 61](#_Toc84500772)

[F1.1.1风洞本体 61](#_Toc84500773)

[F1.1.2运行控制系统 64](#_Toc84500774)

[F1.1.3加热器系统 65](#_Toc84500775)

[**附件2案例编辑模板** 67](#_Toc84500776)

[**附件3知识管理案例** 69](#_Toc84500777)

[F3.1产生式规则的故障诊断专家系统应用案例 69](#_Toc84500778)

[F3.2基于AI的故障诊断专家系统应用案例 76](#_Toc84500779)

1前言

风洞作为空气动力装备的重要试验设施和设备，是发展高新飞行技术和先进飞行器的重要战略装备和基础。随着航空航天行业的飞速发展，风洞设备的规模和数量不断增加，风洞试验任务量也呈现出迅猛增长的趋势，所模拟的试验工况也变得日渐严苛和恶劣，导致风洞设备的可靠性和安全性面临极大的考验。为了保障风洞设备在运行过程中保持良好的性能状态，防止故障导致的意外事件发生，降低风洞设备的运维成本，就必须对风洞设备开展预测与健康管理（Prognostics and Health Management, PHM）工作。

在风洞运行过程中，风洞设备操作人员、试验操作人员和工程师等专家，将会依据设备和试验的具体情况，分析设备的退化和故障原因和机理，产生大量定制化的经验、认知和判断等专家知识。在PHM工作开展过程中，专家知识的加入使得风洞设备PHM模型更加定制化、个性化，可以在一定程度上提高故障诊断、故障预测、健康状态评估等任务的准确度，为风洞设备的健康管理提供着宝贵且有价值的知识。

随着风洞设备的建设发展，风洞试验任务的日趋增长，试验装备的维护难度与日俱增，为保障试验装备正常运行所花费的维修费用和时间也越来越多。过去参试装备的故障监测和诊断主要靠人工方式，效率和数据利用率低。这一过程所产生大量专家经验存在诸多问题。一方面，已经积累的专家知识，形式多样（如条目化规则知识、故障树、案例库等），难以检索和推广利用；另一方面，专家知识管理手段已经不能适应当前风洞平台规范化、标准化发展的要求。因此，将科技管理人员在工作过程中积累的经验收集、存储与使用，构建风洞健康管理知识平台，是提高风洞健康一体化管理水平的关键。

为了风洞健康管理知识平台的顺利构建，本咨询报告从知识管理的角度梳理相关业务要求、标准和技术。针对基地风洞运维专家知识的特点和应用需求，提出基地风洞运维专家知识描述、存储与应用规范化要求，制定知识信息录入模板、规则模板等，以便建立专家知识规则映射表，为知识梳理、平台搭建统一管理的专家知识库提供方法指导和规范要求。

本报告分为5章，对风洞运维专家知识规则化研究的内容进行阐述，各章节主要内容如下：

第一章是前言，首先，介绍基地知识管理现状及开展知识规范化管理的必要性。其次，对基地专家知识管理问题进行剖析，进而对知识规范治理的需求进行分析。

第二章是知识分类及描述方法选用规范，首先，根据风洞健康管理一体化平台的需要和标准，将知识划分为不同类型，其次，介绍了5种主流的知识规则化描述方法，最后，以知识分类为依据规范了知识规则化描述方法的选用。

第三章是专家知识抽取与规则化描述，以自动化程度为切入点，分析了从不同来源、结构的数据中进行知识抽取的3种方式。其次，以人工抽取方式为主，给出了知识规则化的录入模板，对于规则化结果，阐述了图知识和表知识的存储规范。

第四章是知识应用规范，首先，规范了知识库设计的基本内容，随后，提出了知识应用的评价准则，最后，阐述了知识维护的必要的规范流程及内容。

第五章是初步应用建议，根据基地当前知识管理情况，给出了专家知识抽取、表达、存储、维护的推荐方法。

2知识分类及描述方法选用规范

风洞设备健康管理过程中所产生和使用的文档虽得以完整存档，但存在类型单一、知识内容不具体、维度不明确、现行管理不方便应用等问题，缺乏高效的分类和结构化管理。构建风洞设备健康管理知识库的首要问题是：基地试验设备健康管理形成的知识文档结构和成文模式，并不能遵循统一的知识规则化描述方式。因此，需要针对不同知识类型的特点选择合适的知识规则化录入模板。首先，依据四条分类准则划分了知识类型并分析各类知识的特点与判别依据。其次，介绍了5种主流的知识规则化描述方法，最后，以知识分类为依据规范了知识规则化描述方法的选用。

2.1知识分类规范

为了使操作人员能够系统地划分知识类型并选择知识规则化工具，本节梳理了知识分类准则并规范了划分依据，如表2-1所示。随后，根据各类知识特点推荐了适用的知识描述方法。

知识管理方法和使用目的等的不同产生了多种知识分类依据，根据风洞健康管理一体化平台特定的需要和标准，通过比较，把基地的相关知识按照可呈现程度、形成方式、描述内容和功能用途四方面的依据划分成为不同类别的知识，以此显示其在知识整体中的应有位置和相互关系。知识管理的对象是显性知识，因此需要将隐性知识转化为显性知识。通过文字语言、数字符号等将将个体、团队的隐性知识表达出来，就会变为显性知识，主要包括撰写心得、团体总结等方式。

2.2知识规则化描述方法及选用规范

通常专家经验规则所包含的信息可供人清晰辨别，但对于机器而言则过于模糊、抽象，不利于智能化管理。这是因为人类可以接受对事物的模糊和不确定的定义，而基于冯诺依曼构架的计算机在解决问题时需要确定的问题描述及确定的问题解决方案。人类与计算机之间存在着模糊与确定的鸿沟，因此在建立健康管理一体化平台知识库前，需要考虑将人类专家经验规则如何规则化、确定化，使其能转换为计算机可以自动理解判断的知识规则。

2.2.1产生式规则

产生式规则表示法是一种自然的、逻辑的知识表示方法，符合人类关于事物的思维模式。每一条规则具有表示形式一致性，且方便建立规则库对知识进行一

表2-1 知识类型判断依据

|  |  |  |  |
| --- | --- | --- | --- |
| 分类准则 | 知识类别 | 描述与特性 | 判别依据 |
| 可呈现程度 | 显性知识 | 显性知识是指可以用文字、数字等清楚地表达出来的知识，具有可描述性、有除大脑之外的物质载体、具有普遍性和传播性 | 以书籍、设计文件、图纸、磁盘等载体记录的文字、语言、数据、图形、图像、视频等 |
| 隐性知识 | 隐性知识是指个性化，只可意会，难以形式化、公式化的知识，具有个体差异性，不确定性 | 可意会而不可言传的经验、诀窍、技巧 |
| 形成方式 | 经验知识 | 经验知识来源于个人或团队在解决实际问题过程中的记忆和直接感受，难以被编码或作为有形载体呈现。具有主观性、使用局限性 | 领域内的专家的经验知识 |
| 理论知识 | 理论知识是实践获得的感性材料通过归纳整理、抽象而形成的概念与公理，具有客观性、抽象性 | 经过反复验证形成的一些概念、判断和推论，如企业标准 |
| 描述内容 | 过程知识 | 过程性知识指规律性的知识，主要由该领域内的规则、定律和经验构成 | 知识的来源是试验过程、工艺技术等，有确切的公式或步骤来表达 |
| 事实知识 | 事实性知识也被称为叙述性知识，主要是用来描述问题或事物的概念、属性、结构等 | 知识的来源是书本、文献等方面，不需要启发式思考 |
| 功能用途 | 规则知识 | 高度精简地、处理成条目的知识 | 由前提条件和结论两部分组成 |
| 故障树知识 | 用事件符号、逻辑门符号和转移符号描述系统中各种事件之间的因果关系 | 按照GJB/Z 768A-1998构造的故障树 |
| 预案知识 | 根据评估分析或经验，对潜在的或可能发生的突发事件的类别和影响程度而事先制定的处置方案 | 历史出现的各类故障情况所记录的归零报告、故障记录、试验记录、处置程序等 |

致性和完整性查询，能够有效表达过程性知识。规则的完备性、独立性和完整性需要考虑。产生式表示是目前专家系统中使用最广泛的知识表示法。

（1）基本形式

此方法三维基本逻辑结构表现为



可以用以下公式表示：



表示第*k*条规则，表示其结论。在产生式规则中，存在条件的复用和结论与条件关联的情况，即



或



（2）不确定下形式

不确定推理时，产生式规则可以非常方便的结合置信度因子并构成C-F模型。



E表示规则的条件前提或证据；H表示规则的结论；E、H可以是单一或组合的；CF(H,E)是知识的置信度，在[1,1]内取值，称为可信度因子CF，也称规则强度，表示专家对结论H在条件E成立的情况下的信任程度。

2.2.2知识图谱

知识图谱（Knowledge Graph）是Google公司于2012年在知识图谱（Semantic Network）的基础上提出的概念，其主要目的是为了增强搜索引擎的能力，以提高用户搜索质量。知识图谱是一种表达能力强的知识表示方法。不同于产生式表示法对事物因果关系的描述，知识图谱的实质与内涵知识图谱是一种将显性知识和隐性知识融通的途径，其实质是用简单明了的可视化技术将个人拥有的各种资料、信息通过组合形式描述、演变成更具价值的知识体系的一种方法。其中的“图”指的是图示，“谱”指的是系统，“图”与“谱”合一则是知识空间与时间动态变化的统一表述。

知识图谱是一张描述领域知识的带标记的有向图，节点表示各种事物、概念等对象，边表示节点之间的语义关系。节点带有若干属性，一般用框架或元组表示。节点还可以是语义子网络，从而形成多层语义嵌套网络。

从结构上来说，知识图谱是由基本的语义单元组成。一般来说，这些基本的语义单元主要可以称为知识图谱中的语义基元。在表示上，知识图谱的基本网元结构是三元组结构为（头节点，边，尾节点）。对于简单的语义关系，通过直接使用语义基元即可表示出来，如图2-1所示

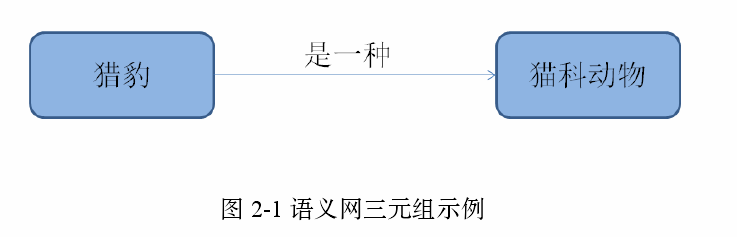


图2-1 语义网三元组示例

对于复杂的语义关系，需要多个语义基元组合使用，甚至需要增加中间节点，如“当河流水位超过警戒线且持续降雨，或者连续降雨天数大于５天时，发出警报”这种既包含“与”关系和“或”关系的复杂语义，需要对语义进行使用多个语义基元和增加中间节点来表示。

知识图谱基本关系主要包括以下类型。

属性类型：Have关系，表示一个节点是另外一个节点的属性。

类属类型：ISA(is-a）关系，表示一个节点是另外一个节点的实例或是子类。

聚类类型：Part-of关系，表示一个节点是另外一个节点的某个部分。

时间序列类型：After/Before/At关系，表示节点间的次序关系。

位置类型：Located-On/At/Under关系，表示节点间的空间位置关系。

组成类型：Composed-of表示节点由另外一些节点组成。

因果类型：If-then表示节点间的因果逻辑关系。

相似类型：Similar-to表示一个节点与另外一个节点相似。

2.2.3框架法

框架表示法是一种结构化的知识表示方法，善于描述结构性的知识。框架之间的嵌套构成框架网络式的框架更能表现事物之间的复杂关系。其继承性消除了知识表达的冗余描述，较好的保证了知识的一致性。框架的结构能够适应多种推理方法，推理机制可以灵活制定。但是，对于过程性知识的描述能力较低，通常会将框架表示法和产生式表示法结合使用。框架允许结点有结构从而增加第三维，而语义网本质是一种二维的表示方法。框架的记录类型与记录值对应的术语是“槽”和“槽值”。通过在槽值中再次使用框架来继承，可以建立更加强大的知识表示系统。框架系统可以很好的描述诸如汽车这样的机械装置结构。

（1）框架法的结构

框架表示法主要用以表示结构固定的数据，框架的顶层是固定的，表示某个固定的概念或事件。下层由多个“槽”组成，“槽”又可以细分为若干个“侧面”，每个“侧面”由“侧面名”和“侧面值”组成。其中，“槽”用以表示某一类型或某一方面的属性知识，“侧面”表示某一属性的具体名称和取值。一个知识表达系统可以由多个“框架”组成，一个“框架”可以由多个“槽”组成，一个“槽”又可以由多个“侧面”组成，不同的“框架”、“槽”和“侧面”以不同的“框架名”、“槽名”和“侧面名”区分。框架结构的一般表达式如表2-2所示。

表2-2 框架法结构

|  |  |
| --- | --- |
| <框架名> |  |
| 槽名1： | 侧面名11：侧面值11 |
|  | 侧面名12：侧面值12 |
|  | … |
|  | 侧面名1*n*：侧面值1*n* |
| 槽名*k*： | 侧面名*k*1：侧面值*k*1 |
|  | 侧面名*k*2：侧面值*k*2 |
|  | … |
|  | 侧面名*km*：侧面值*km* |

（2）相关说明

①框架名的值允许带有参数。此时，当另一个框架调用它时需要提供相应的实在参数。

②当槽值或侧面值是一个过程时，它既可以是一个明确表示出来的<动作>串，也可以是对主语言的某个过程的调用，从而可将过程性知识表示出来。

③当槽值或侧面值是谓词时，其真值由当时谓词中变元的取值决定。

④槽值或侧面值为空时，表示该值等待以后填入，当时还不能确定。

⑤约束条件是任选的，当不指出约束条件时，表示没有约束。

由上述表示形式可以看出，一个框架可以包含任意有限个数目的槽，一个槽可以包含任意有限个数目的侧面，一个侧面可以包含任意有限个数目的侧面值。一个槽可以分为若干个侧面，也可部分侧面，视其描述的属性而定。另外，槽值或侧面值既可以是数值、字符串、布尔值，也可以是一个在满足某个给定条件时要执行的动作或过程，特别是它还可以是另一个框架的名字，从而实现一个框架对另一个框架的调用，表示出框架之间的横向联系。

2.2.4谓词逻辑

谓词逻辑是一种以对象和关系为基本元素将相关知识利用符号形式表达的知识表达方法。是人工智能领域应用较早、较广泛，也是较成功的一种知识表示模式，其表达的知识主要是浅层知识。

在谓词逻辑中，命题是用谓词表示的，一个谓词可分为谓词名和个体两个部分。个体表示某个独立存在的事物或者某个抽象的概念：谓词名用于刻画个体的性质、状态或个体间的关系。

谓词的一般形式是



其中是谓词名，通常用大写英文字母表示；是个体，通常用小写英文字母表示。

把自然语言命题用谓词逻辑表达出来需要以下三个步骤：

（1）将一个原子命题分解为个体词和谓词两个部分，用来代替个体变元，用等来代表个体常项，用大写字母等来代表谓词变元。

（2）找出原子命题中所勾含的量词，表示“所有"，表示“存在”。

（3）运用人工符号来表示原子命题中个体词与谓词的关系及多个个体词或谓词间的复合关系。

（4）构造出该命题所对应的形式化的表达式公式。

2.2.5基于人工智能的方法

知识库和自然语言处理的各个任务都需要对符号或者文本进行表示和处理，并且在此基础上进行语义解析、语义计算、知识推理。然而，文本和符号在引用过程中具有歧义性和动态性，这为语义解析带来了困难，而且传统的文本和符号的离散表示也难以用来进行大规模的语义计算和知识推理，也很难与上层任务建立统一的模型。近年来，深度学习的蓬勃发展为上述问题的带来了改观。

神经网络的知识表达与传统基于规则诊断的专家系统的规则知识表达有着本质的区别，神经网络知识的表示是利用网络自身的分布式连接机制对知识进行隐式表示。知识表示不再是独立的一条条的规则，而是网络的拓扑结构和分布于整个网络中的权和阈值。神经网络专家系统的单个神经元知识表示如图2-2所示。图中X表示来自其它神经元知识的激励信号，W表示神经元之间的连接权值，表示神经元阈值，y表示神经元的信号输出。



图2-2 神经元知识表达

具体地，首先，对知识库做知识表示，将实体和关系嵌入到低维空间。然后，采用卷积神经网络从数据中识别出实体，并表示成结构化的向量形式。接着，采用LSTM等循环神经网络从问题中抽取关系，并表示成结构化的向量形式。获取实体表示和关系表示之后，匹配（链接）到知识库中的最优实体关系对，并基于规则对候选答案进行排序获取最终答案。

2.2.6知识规则化描述方法选用规范

知识表示方法的选择应考虑以下4点因素：①能够充分表示相关的领域知识；②有利于对知识的利用；③便于知识的组织、维护和管理；④便于理解和实现。以这4点因素为综合考虑依据，结合研究目的，构成本研究选择知识描述方式的基本思想。知识规则化描述方法对比如表2-3所示。各类知识的规则化描述方法选用规范如表2-4所示。

2.3知识分类示例

由于非结构文本中影响语义、语用的变量众多，难以处理，自然语言处理在语义、语用方面应用尚不成熟。因此，录入知识时人工判断更为可靠。本节为知识分类规范具体示例，如表2-5所示。

表2-3 知识描述方法特点总结与选用规范

|  |  |  |
| --- | --- | --- |
| 表示方法 | 优点 | 缺点 |
| 产生式表示法 | ①表示形式与接近人类思维模式，便于理解  ②表示的格式固定，形式单一，规则间相互独立  ③可以表示确定和不确定性知识 | ①推理效率低下  ③缺乏灵活性：规则之间不能直接调用，因此较难的表示具有结构关系的知识 |
| 知识图谱表示法 | ①以明确、简洁的方式表示各节点之间的联系  ②着重强调事物间的语义联系，有关事物或概念的相关事实可由它们直接相连的节点推导出来。  ③具有广泛的表示范围和强大的表示能力 | ①逻辑上是不充分的，不能保证网络操作所得推论的严谨性和有效性  ②一旦节点个数过多，推理就难以进行  ③不便于表达判断性知识与深层知识 |
| 框架表示法 | ①人类的思维和问题求解过程相似  ②框架结构表达能力强，层次结果丰富，提供有效地组织知识的手段  ③可以利用过去获得的知识对未来的情况进行预测 | ①缺乏形式理论，没有明确的推理机制保证问题求解的可行性和推理过程的严谨性  ②不善于表示过程性、不确定性知识 |
| 谓语逻辑表示 | ①该方法对简单说明构造复杂事物的方法有明确、统一的规定，有效地分离了知识和处理知识的程序  ③具备完备的推理算法  ④是一种形式推理，不依赖于任何具体领域，具有较好的通用性 | ①难以表示过程和启发式知识  ②由于是弱证明过程，当事实的数目增大时，在证明过程中可能产生组合爆炸  ③表示内容和推理过程分离，推理按形式逻辑进行，内容包含的大量信息被抛弃 |
| AI表示 | 对语义解析、语义计算、知识推理有较好支持 | 规则之间的关系是隐式的 |

表2-4 规则化描述方法选用规范

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 分类依据 | 知识类型 | 特征或需求 | 适用方法 | 适用范围 | 功能边界 |
| 可呈现程度 | 显性知识 | 可描述性、有除大脑之外的物质载体、具有普遍性和传播性 | 全部适用 | 知识规则化的对象是显性知识 | / |
| 隐性知识 | 无法用语言和符号表示，需要显性化 | 全部不适用 | / | / |
| 形成方式 | 经验知识 | 主观性、使用局限性 | 产生式表示法 | 对象与产生式形式相近 | 需要以条件指向结果 |
| 理论知识 | 客观性、抽象性 | 谓词逻辑 | 描述对象是强有力的逻辑关系事物 | 需要精确、严密表达 |
| 描述内容 | 过程知识 | 描述步骤、流程 | 产生式表示法 | 对象需要以条件指向结果 | 需要以条件指向结果 |
| AI | 对象涉及解析语义 | 需要文本处理和推理 |
| 事实知识 | 内部结构关系及知识之间的特殊关系表示出来 | 框架法 | 对象内部有重要结构关系 | 保证结构、关系清晰 |
| 谓词逻辑 | 描述对象是强有力的逻辑关系事物 | 侧重刻画关系 |
| 知识图谱 | 涉及关系表达 | 需刻画结构关系 |
| 功能用途 | 规则知识 | 具有前提部分和结论部分 | 产生式表示法 | 对象与产生式形式相近 | 需要以条件指向结果 |
| 故障树知识 | 将故障树各推理节点的故障现象与实际测试数据、计算公式映射成计算机易于处理的结构 | 框架法 | 对象内部有重要结构关系 | 保证结构、关系清晰 |
| 产生式表示法 | 对象与产生式形式相近 | 需要以条件指向结果 |
| 知识图谱 | 对象的各节点之间的联系 | 需要刻画结构和推理 |
| 预案知识 | 描述案例处置程序、原因、结果、风险分析等内容 | 框架法 | 对象内部有重要结构关系 | 保证结构、关系清晰 |
| AI | 涉及解析语义 | 自动化文本处理 |
| 知识图谱 | 涉及比较复杂推理 | 需要推理 |

表2-5 知识分类示例

|  |  |  |
| --- | --- | --- |
| 源知识 | 符合的判别依据 | 分类结果 |
| Φ1米高超声速低密度风洞预抽真空泵系统主要由ZJQ600气冷罗茨真空泵、H150滑阀式真空泵等组成。泵管路之间均配有溢流阀、返流过滤器、列管式冷却器及温度控制仪表，使该机组可在大气压下直接启动，在较高的进口压力下长期可靠运行。 | 知识的来源是书本、文献等方面，不需要启发式思考 | 事实知识 |
| 时间：2020年2月25日，发现人：李杰、龙正义  故障及维修经过：  按下启动键后机组发出嗡嗡的声音,经仔细检查，发现一根电线熔断，导致电机缺相，并且相邻的电线都有不同程度烧损。由于疫情期间需临急恢复试验，联系省安装公司相关人员前来维修，清洗了前级泵，更换了电线、接触器、热继电器及电流表等元器件。  原因分析：  前级泵内油垢积累过多，导致泵在运行过程中负载过大，加之长时间运行，过载过热严重，导致电线熔断。  经验教训：  1、每次运行前检查真空泵油油位是否在安全位置；  2、定期对真空泵进行清洗保养；  3、每次运行前对电气柜里各类元器件进行检查。 | 历史出现的各类故障情况所记录的归零报告、故障记录、试验记录、处置程序等 | 预案知识 |
| 若20MW送进系统不能运行自动步骤，尝试重新启动PLC系统 | 领域内的专家的经验知识 | 经验知识 |
| 富氧空气管在试验设备工作时尾端加速度达到30g；中部加速度达到1g；前端加速度达到1.5g；且富氧空气管为静态部件。因此选取振动传感器量程应大于测定值1.5倍 | 知识的来源是试验过程、工艺技术等，有确切的公式或步骤来表达 | 过程知识 |
| 喷管在试验设备试验过程中振动能量分布在2000~2300Hz和3500~3700Hz附近，则三轴振动传感器频响范围为0.5~7000Hz，量程+100g | 经过反复验证形成的一些概念、判断和推论，如企业标准 | 理论知识 |

3专家知识抽取与规则化描述

在上一章中将知识划分为不同类型，根据知识类型规范了知识规则化描述方法的选用。本章介绍确定知识类型后，如何从源知识中抽取所需知识并使用规则化方法存入数据库。首先，根据知识抽取的自动化程度，将抽取分为人工、半自动化和自动化3种方式，介绍了对应的主流方法、工具。其次，以人工抽取方式为主，给出了知识规则化的录入模板，对于规则化结果，阐述了图知识和表知识的存储规范。

3.1知识抽取方式

知识抽取（Knowledge Extraction）研究如何根据给定本体从无语义标注的信息中识别并抽取与本体匹配的事实知识，即从无语义信息的文档内容中抽取与本体匹配的事实知识，进而实现数据充分、有效的利用。该技术可以抽取出事实知识用于构建基于知识的服务。信息抽取的完整度、准确度直接显性影响后续知识图谱构建步骤的质量和效率以及最终知识图谱的质量。

可以将知识抽取描述为这样的一个过程：首先，第一个阶段对大量孤立、模糊、复杂的动态非结构化数据进行初步处理和计算；然后，第二个阶段对数据进行深层语义分析、用户隐私保护问题分析以及应用领域知识的结合分析；最后，第三个阶段选择合适的挖掘算法和抽取技术进行数据抽取和融合。通过将抽取得到的碎片化知识存入知识库的数据层和模式层，我们最终可以对数据形成本体化表达。

抽取技术按照抽取过程可以分为实体抽取（Entity Extraction）、关系抽取（Relation Extraction）、属性抽取（Attribute Extraction）以及实体链接（Entity Linking）等。其中，实体抽取用于发现文本或者网页中的命名实体，并将其加入现有知识库中。关系抽取用于自动抽取实体之间存在的语义关系。属性抽取属于一种特殊的关系抽取。信息抽取的目标是自动化知识获取，即实现自动地从异构数据源中抽取实体、关系、属性等信息进而得到候选知识单元。

从知识抽取过程的自动化程度来看，知识抽取又可以分为人工抽取、半自动抽取和自动抽取三种方式。自动化程度的不同反映了知识抽取系统所具有的推理能力的不同，以下将对这三种抽取方式进行介绍。

3.1.1人工抽取

人工抽取又称为基于规则的方法，主要是由知识工程师与领域专家共同工作、交流来提炼知识、对所需要的知识内容及类型进行定义，即人工编写规则。在提炼出规则之后，知识工程师就可以根据提炼出的规则对大量的结构化数据、半结构化数据和非结构化数据进行处理，最终将其转变为统一格式的结构化数据。最终的结构化数据可以以图或表的形式方便地存储。这种方式首先要先构建大量的实体及关系抽取规则，然后再将规则与文本字符串进行匹配，识别出命名实体及它们之间的关系。

在知识工程师与领域专家交流协作时，可以通过下面几种方法获取到需要的知识信息并编写规则：

（1）面谈法：与专家进行面对面交谈是一种广泛使用的知识获取方式，但是不拘形式的会谈不易获得详细的知识，因此，可采用专题面谈的形式，即向专家提出事先拟好的问题，由专家任意回答。问题可大致分为两种类型，即“在XXX情况下将怎样处理”和“为什么这样做”等。

（2）模拟法：模拟法可分为静态模拟和动态模拟。静态模拟是提出某一实例的情况，请专家谈其求解过程。这一方法肯定会得到一套资料，说明其求解过程所用的知识和步骤。它的优点是，因系静态条件下进行，可以集中到我们最感兴趣的方面讨论。它的缺点是，工作压力与时间压力与实际情况不同，且可能忽略某些细节。动态模拟是在专家处理某真实问题时，知识工程师观察并录下其实际求解步骤，然后再进行分析。这种方法的优点是能够观察到专家在自然状态下的工作过程，缺点是较为费时费力。

（3）口语记录分析：当专家解决问题时，让专家自己叙述自己想些什么，并加以笔录或录音，然后对记录加以分析，即对专家的思维活动做出叙述性的记录。在口语记录分析中主要处理两类问题：一种是分析表达明显的部分，即从口语记录明显看出专家思维过程的部分；一种是隐含表达的部分，即对专家没有明确表达的思维过程，需要分析者根据记录加以推断，他的某些语句时基于什么背景得出的。只要专家充分合作，知识工程师有足够的经验和知识，口语记录分析就是一种可靠的知识获取方法。当然在关键地方还可以使用面谈法加以考核，以防失真。

三种方法的特点及其适用范围见下表3-1：

表3-1 获取知识信息的三种方法对比

|  |  |  |
| --- | --- | --- |
| 方式 | 特点 | 适用范围 |
| 面谈法 | 知识工程师与专家面对面交流以获取想要了解的信息 | 适用于知识工程师与专家一起工作或不受时间和空间的限制时 |
| 模拟法 | 知识工程师提出问题让专家求解或是知识工程师观察专家的工作过程 | 适用于知识工程师能够贴近真实工作情况时 |
| 口语记录分析 | 知识工程师记录下专家解决问题时的思路 | 知识工程师有充足的经验和知识，专家充分合作 |

在构建出规则之后，可以按照以下步骤进行知识抽取：

（1）对构建出的规则进行解析；

（2）对句子进行分词、词性标注、命名实体识别、依存分析等处理，并构建出依存树；

（3）在句子依存语法树上匹配规则，每匹配一条规则就生成一个三元组；

（4）根据扩展规则对抽取到的三元组进行扩展；

（5）对三元组实体和触发词进一步抽取出各个实体之间的关系。

其流程图如下图3-1所示：



图3-1 人工知识抽取流程图

早期的知识抽取大都采用这种方式，在小数据集上可以达到很高的准确率和召回率，所得到的知识大多可以直接用于解决问题，但是不适合用于大数据集。对于风洞一体化平台中的数据即可采用此种抽取方式，能够取得较好的效果。

3.1.2半自动抽取

半自动知识抽取指利用一些知识获取工具来辅助知识工程师把知识原材料或专家描述的知识内容经过识别、理解、筛选、格式化后以一定的形式存入知识库中。其主要包括以下6个步骤，流程图如下图3-2所示：

（1）用一组信息模式描述感兴趣的知识，如竞争情报采集与监测系统中常用到的“推出<公司，产品>”；

（2）对数据进行预处理，包括格式清洗与转换；

（3）对数据进行词法分析、浅层句法分析以及简单的语义分析，识别名词短语并标注语义信息；

（4）使用模式匹配方法实现事件模板的构造，建立实体之间的关系；

（5）采用语段分析技术实现句子相关性分析，进行上下文关联、指代消解、引用等分析和推理，构造一个完整的实体事件；

（6）格式化分析结果，把抽取的信息按预定义的模板输出。

形状

低可信度描述已自动生成

图3-2 半自动知识抽取流程图

在半自动知识抽取中，需要要对海量的数据进行预处理，这是一项人力所不能完成的任务，因此需要依靠计算机的协助。自从科学计量学分析引入辅助可视化以来，科学知识图谱领域也诞生了一批有价值的科学知识图谱工具。表3-2列出了一些常用的知识图谱辅助工具。

表3-2 常用知识图谱辅助工具

|  |  |  |
| --- | --- | --- |
| 工具名称 | 特点 | 适用范围 |
| BibExcel | 可对数据格式转换及去噪，并进行BCAD、CAAA、CAAC、ACA、DCA、CWA等分析 | 适合用于简单的计量分析以及为一些可视化工具提供辅助， 如统计、构建共现矩阵等 |
| CiteSpace | 可对数据进行去重和时间切片，并进行BCAD、CAAA、CCAA、CAAI、ACA、DCA、JCA等分析 | 适合用于时间序列分析、突发监测 |
| SPSS | 可用于相关分析、因子分析、主成分分析、多维尺度分析、聚类分析等 | 适合用于分析偏小、分布均匀的数据集 |
| TDA | 可用于聚类分析、关联分析、时间序列、突发监测等 | 适合用于数据清洗、构建矩阵，为SPSS和Ucinet等一些可视化工具提供支持 |
| Ucinet | 可用于中心性分析、子群分析、角色分析、聚类分析等 | 适合用于分析较大的数据集、分析核心关键词 |

使用相关工具来生成文献分析的可视化图谱固然重要，但认识所分析的数据内容及其结构是我们进行图谱绘制的前提。因此，还可以使用Notepad++和Sublime Text这两个工具快速打开所分析文本，并查看其内容和结构。

采用半自动抽取时可以把文档当作字符串进行训练，如果文档是比较结构化的还可以提取出结构化文档中的树结构去训练。

使用半自动抽取在学习到抽取规则后可以有较好的效果，但是手工标注训练数据代价高昂，不仅要标注一定数量的训练数据，而且要选择适当的训练数据，会比较繁琐。因此这种方式只适合在数据量较小或数据格式统一时使用，在数据量较大且数据格式不统一时较为麻烦。对于风洞一体化平台，如果要使用半自动抽取方式需要预先对数据进行处理，较为麻烦且易出错，因此可避开这种方式。

3.1.3自动抽取

自动抽取利用数据挖掘、机器学习、深度学习等人工智能技术通过对应用实例与实际问题进行建模从而发现一些专家尚未形式化甚至未发现的新知识和新规律，从信息源中自动提取出实体、属性和关系。

近年来，基于预训练模型通过微调进行学习的方法在自然语言处理领域取得了巨大成功，Google的BERT模型就是其中的重要代表。

BERT（Bidirectional Encoder Representations from Transformers）是一种基于转换器（Transformer）的双向编码表示模型，它的拓扑结构是一个多层的双向转换器网络。BERT模型是基于微调学习的典型应用，也就是说他的构建包含有预训练和微调两个步骤。

首先在预训练阶段，对大量不同训练任务的未标记语料数据进行训练，将语料中的知识迁移进了预训练模型的文本嵌入（Embedding）中。这样，在微调阶段，只需要在神经网络中增加一个额外的输出层，就可以对预训练模型进行调整了。具体而言，微调就是用预训练参数初始化BERT模型，然后，使用来自下游任务的标记数据对模型进行微调。

如图3-3所示是使用BERT从保险文档中抽取知识点的模型流程。



图3-3 使用BERT从保险文档中抽取知识点

在自动抽取知识的时候，首先要使用大量的训练数据即训练语料对机器学习或深度学习模型进行训练，在训练完成之后再去使用训练好的模型抽取新的知识。

这种方式构建出模型的代价较小，鲁棒性更好，但是训练语料的数量和质量直接影响着这种方法的抽取性能，知识畸变（言语描述和具体行动间形成的错位现象）也影响着知识获取的精准性，自动抽取方式还未成熟。基于以上种种原因，自动抽取方式还不适合实际应用。

以上三种抽取方式的对比见表3-3。

表3-3 三种抽取方式对比

|  |  |  |  |
| --- | --- | --- | --- |
| 抽取方式 | 方式 | 特点 | 适用条件 |
| 人工抽取 | 人工编写规则，手动分析、整理 | 不需要训练模型，简单方便 | 适用于小数据集 |
| 半自动抽取 | 手工标注数据后使用机器学习方法学习规则 | 节省人力 | 适用于中等规模数据集，能够手工标注数据且选择合适的训练数据 |
| 自动抽取 | 使用人工智能技术进行建模从而抽取知识 | 自动化程度高、建模代价较小、鲁棒性高但是技术还不成熟 | 适用于大数据集且自动化程度要求较高时 |

3.2知识规则化

本小节对5种知识规则化方法的流程进行规范，以建立知识录入模板。

3.2.1产生式规则

作为专家系统中应用最广泛的方法，产生式规则简单易行便于理解，适合经验类知识规则化，但不能表示结构性知识。该方法应用流程主要步骤如图3-4所示。



图3-4 产生式规则的应用流程

（1）知识获取：知识获取是指从一个或多个知识源发现、吸取、构造和组织知识，使之形成系统知识库的过程。一种途径是由知识工程师通过和领域专家交谈，以及阅读、分析各种资料得到关于领域的各种知识，然后把知识输入到计算机中；另一种途径是通过机器学习，从处理问题的过程中获得。

（2）知识库存储结构设计：产生式规则知识库由字典库和规则库2部分组成。字典库用于存放现象和结论；规则库包括规则前件库和规则后件库，结论有可能仍然是前提，所以规则前件用于存放规则的前提条件和前提规则号，规则后件库用于存放规则的结论部分和后推规则号。结论允许包含该结论的置信度，用于在结论中进行有效筛选

（3）产生式规则到知识库的转化：产生式的知识表示与关系数据库记录的表示有着良好的接口，因此较多采用关系数据库开发。关系数据库的数据结构是一个二维表，它是以二维表结构来描述客观世界的实体及其联系。表的每一行对应一个元组，每一列对应一个属性数据按属性分解，按元组存储。用数据库表示产生式规则前，需进行处理：利用逻辑关系“或”的等价转换关系将一条规则中前提的混合逻辑关系转换为只含逻辑关系“与”的单纯关系。逻辑关系“或”的等价转换关系如下面的公式：



基于上述步骤，制定产生式规则下的知识录入模板如表3-4所示。

表3-4 产生式规则录入模板

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 规则号 | 条件数 | 条件1 | … | 条件*m* | 结论 |
| Rk001 | *m* | E11 | … | E1m | C |
| Rk002 | *m* | E21 | … | E2m | C |
| … | … | … | … | … | … |
| Rk00n | *m* | En1 | … | Enm | C |

需要注意的是，产生式规则的录入需满足条件最小化单元原则，即导致结论成立的条件应不可继续向下分解。

表3-5是一个产生式规则的示例，表示：如果“真空泵腔内没有适当油量且油路不通”或“真空泵进气管过滤网堵塞”，那么“真空泵真空度低”。

表3-5 产生式规则示例

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 规则号 | 条件数 | 条件1 | 条件2 | 结论 |
| R1001 | 2 | 真空泵腔内没有适当油量 | 油路不通 | 真空泵真空度低 |
| R1002 | 1 | 真空泵进气管过滤网堵塞 |  | 真空泵真空度低 |

常用开发工具有：

（1）用于专家系统开发的通用程序设计语言的主要代表有C++、C#、PASCAL、ADA等。

（2）数据库端：SQL Sever、MySQL、Access等。

（3）传输接口：ADO、ODBC等。

（4）用户界面：B/S架构，C/S架构。

3.2.2知识图谱

知识图谱是一系列可以用来展示知识的发展和结构关系的图形，它充分釆用可视化的技术，不仅能够对知识资源和载体进行描述，同时还可以对知识以及知识之间的联系进行分析和描绘。当前，知识图谱主要应用于科学研宄领域。该方法的应用流程主要步骤如图3-5所示。



图3-5 知识图谱的应用流程

（1）知识抽取

知识抽取，即从不同来源、不同结构的数据中进行知识提取，形成知识（结构化数据）存入到知识图谱，知识抽取的内容包括实体抽取，关系抽取和属性抽取。

①实体抽取

实体抽取又称命名实体识别，其目的是对于给定数据，从其中识别并抽取出具体的人名、地名、机构名、时间等实体并划分到对应的类别中。目前对于自然语言文本的命名实体识别方法大概分成以下两类：基于词典和命名规则的命名实体识别方法、基于机器学习的命名实体识别方法。

a.基于词典和命名规则的命名实体识别方法

基于词典和命名规则的方法主要是通过人工制定抽取模板，以模式和字符串匹配为主要手段，依赖已有的词典知识库，从非结构化文本数据中识别出实体，并划分到对应的类别中。这种方法需要语言学专家和计算机专家相互配合构建，很费时费力。

b.基于机器学习的命名实体识别方法

基于机器学习的命名实体识别方法其核心是分类问题。其中分类处理又包含两种：一种是根据给定的实体类别来识别和划分实体边界，然后进行实体分类，另一种是设定若干个待标注的类别标签，然后自动化对文本内容进行标签标注，最后将标注的标签整合，最终构成命名实体。目前标注效果较好的模型有隐马尔可夫模型、长短期记忆网络、卷积神经网络模型等。然而，基于机器学习方法无法捕获实体的语义特征。深度学习方法将识别过程转换为序列标注问题，通过设定标注规则为每个词打上类别标签，从而得到命名实体识别结果。最经典的识别效果比较好的模型是LSTM-CRF。

②关系抽取

关系抽取是指根据命名实体识别出多个实体之间存在的关系，关系抽取的结果是三元组形式，关系抽取的方法可以分为基于模板和基于机器学习的方法。基于模板的方法主要是通过设计语义抽取规则来进行的。基于机器学习的方法主要包括有监督的关系抽取方法和弱监督的关系抽取方法。

③属性抽取

属性抽取是指通过属性抽取技术抽取出实体的属性值，通过实体的属性值集合来对该实体进行整体描述。在属性抽取中三元组表示形式为（实体，属性，属性值），可以将关系抽取技术用于属性抽取。

（2）知识融合

知识融合(Knowledge Fusion)即把多源异构的知识经过实体对齐、实体消歧等技术进行融合成新的知识，并获取价值。实体对齐方法中，具有代表性方法有JE、MTransE、ITransE、KDCoE等。

（3）可视化

前端界面将可视化请求发送给后端，后端调用知识库查询接口得到知识数据，并将其转换为前端界面可以识别的包含节点和关系的三元组形式数据集合，并发送到前端界面构建相关图例展示给用户。在进行前端可视化时，通过获取特定数据生成json格式进行处理，之后将数据返回给前端界面。可视化流程如图3-6所示。



图3-6 知识图谱可视化流程图

知识图谱下的知识录入模板应是一系列三元组（实体，关系，实体）或（实体，属性，属性值）的组合，下图是一个知识图谱的示例，表示：编号001的电机发出“嗞嗞”声代表该电机已故障。



图3-7 知识图谱示例

常用开发工具有：

（1）知识图谱可视化：RDF、OWL、Turtle等。

（2）存储：图数据库，主流的图数据有Neo4j、HugeGraph、GraphDB等。

3.2.3框架法

框架表示法善于描述结构性的知识，相对于产生式规则，能够提供更加细致、复杂的知识描述。该方法的应用流程主要步骤如图3-8所示。



图3-8 框架法的应用流程

（1）词法分析：针对框架知识库以及问题库，这个阶段主要是对知识进行扫描，识别出各单词以及标点符号，并建立单词链表，为语法分析做准备。

（2）语法分析：主要任务是识别由词法分析给出的单词符号序列在结构上是否符合给定的文法规则。具体方法是扫描词法分析得到的单词链表，判断知识输入格式的正确性，并根据事先定义的一个保留字库识别出知识中的保留字及用户自定义符号。得到框架知识库中框架、槽、侧面的链表结构（分为定义框架和应用框架），并判别槽、侧面的取值是否符合定义要求。

（3）语义分析：主要任务是在语法检查的基础上，判断各个框架之间的关系、槽与框架之间的关系、槽与槽之间的关系以及槽与侧面之间的关系，为推理回答问题做准备。由于槽和侧面的值都可以是另外一个框架，因此在框架展开时会出现无限循环及矛盾等情况。解决这些问题的方法是检测出循环点和矛盾点并终止该过程。

基于上述步骤，制定框架法下的知识录入模板如下：

表3-6 框架法录入模板

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 框架名 | | <框架名> | | | |
| 槽名1 | 侧面名1 | 值1 | 值2 | … | 值 |
| … | … | … | … | … |
| 侧面名 | 值1 | 值2 | … | 值 |
| … | … | … | … | … | … |
| 槽名*n* | 侧面名1 | 值1 | 值2 |  | 值 |
| … | … | … | … | … |
| 侧面名 | 值1 | 值2 | … | 值 |

表3-7是一个框架法的示例，最上层框架是三相异步电动机故障框架，实际上是空框架，只是包含了下层一些子框架的名字。

表3-7 框架法示例

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 框架名 | <三相异步电动机故障> | | | | | | |
| 故障征兆 | 不能转动 | | … | | | 温度突然升高 | |
| 维修建议 | <电动机故障维修建议框架> | | | | | | |
| 父框架 | 空 | | | | | | |
| 子框架名 | <转子故障> | <定子故障> | | <偏心故障> | <轴承故障> | | <绝缘故障> |

常用开发工具有：

（1）使用JSON数据格式更新数据存储；

（2）底层数据管理平台使用MongoDB；

（3）知识创建录入前端使用Jquery框架；

（4）使用AJAX（Asynchronous Javascript And XML）技术传输数据。

3.2.4谓词逻辑

谓词逻辑以对象和关系为基本元素将相关知识利用符号形式进行表达。该方法的应用流程主要步骤如图3-9所示。



图3-9 谓词逻辑法的应用流程

（1）概念层次划分：概念层次的划分，将问题层层向下分层分类。合理的概念层次划分能够使应用中相关事实和规则知识的表达精确高效。

（2）对象类别属性定义：要想实现知识的表达，还需要对上述对象类别的属性进行定义，这些属性中一方面包含例如设备等级，额定容量，设备型号的静态信息，另一方面还包含表示运行状态的动态信息，如设备是否投入，开关的状态等。

（3）对象关系定义：基于对象类别的划分对对象类别间关系进行定义；对象间关系的表示格式定义为<关系主体，关系，关系客体>。

（4）形式化表达：①合理选择谓词排序：自左向右逐一匹配，当遇到无常量赋值满足当前项的变量约束时即返回匹配不成功结果，终止规则剩余部分谓词的匹配赋值。②避免条件重复：规则中，往往一个子条件的不同会导致不同的结论；针对这种情况，若对于每条规则含有大量相同子条件，且分别书写完整的规则，在进行规则匹配时将造成规则相同部分的重复匹配，降低匹配效率，因为规则在匹配时采用的是正向匹配，所以在建立规则库时对于多条规则的相同部分可提取出来，单独列写为一条规则，并将其结论设定为中间结论，再用其替代原先规则间相同的条件部分，便可减少重复部分的多次匹配。③设定规则优先级：由于知识库的不断扩充，新增的规则可能会与原有规则同时匹配成功发生冲突，这时就需要对规则间的优先级进行说明；因此在规则库中增加优先级栏，栏中数据类型为列表，若多条规则满足，从优先级列表的第一个参数开始比较，数值越小优先级越高，若一样则比较第二个参数大小，直至确定优先等级，若列表长度不一样则默认缺省数据为0；这样设计的一个好处时在添加规则时，无需更改原有的规则优先级说明，只需对新添加规则按照优先等级进行列表数据设定即可。

谓词逻辑方法的知识录入模板一般是具体知识及对象而定，举例来说，对于规则“若一条故障线路各端开关均断开，则故障线路已隔离”，可表示为：

∃x.(Line(x)&Fault\_equip(x)&∀y.(Breaker(y)&connect(x,y)&Breaker\_state(y,off))) => Isolated\_equip(x)

上述表达规则中，x、y为变量符号；&和=>为逻辑连接词，&代表逻辑与，=>代表蕴含，表示在符号左侧事实为真的情况下右侧事实一定为真；Line、Breaker表示线路和开关的设备类型，Breaker\_state表示开关的开断属性，off表示断开属性值，这里以常量符号表示；Fault\_equip表示故障设备，Isolated\_equip表示隔离设备。

常用开发工具可使用与产生式规则相同的开发工具，也有项目采用Visual Prolog、MySQL VIP等工具。

3.2.5基于人工智能的描述

基于人工智能的知识描述是隐式的，将要表示的知识的有关信息分布在网络中的权和阈值中。这种分布式的表示方法便于知识库的组织和维护，可以拥有大量的知识。还有如下特点：

①便于后续推理功能开发：神经网络知识库体现在神经元之间的连接强度（权值）上。它是分布式储存的，适合于并行处理。一个节点的信息由多个与它连接的神经元的输入信息以及连接强度合成。

②神经网络有成熟的学习算法：学习算法与采用的模型有关，对于反转传播（BP）模型来说，模型采用误差沿梯度方向下降以及隐含层节点的误差由输出节点误差反向传播的思想进行。通过反复的学习，逐步修正权值，使之适合于给定的样本。

③容错性好：由于信息是分布式储存的，在隔壁单元上即使出错或丢失，所有单元的总体计算结果可能并不改变。

④灵活性好：基于人工智能的方法没有固定不变的模板，可根据特定领域开展训练或迁移学习。

基于人工智能的方法可以看作文本内容的词标注或字标注问题，具体内容包括中文分词、词性标注、命名实体识别、语义角色识别等。具体地，通过给定一段文本，将文本映射到字向量空间，利用神经网络提取字特征，利用卷积神经网络提取句子特征，利用普通神经网络提取类别特征，最后首尾相接的三类特征经过全连接分类层得到字的候选标注。使用人工智能录入知识流程如下：

（1）训练样本

网络的训练样本从知识获取阶段获取到的故障诊断信息实例中抽取，并且通过信息预处理对将结构化、编码化的信息作为网络的训练样本。

（2）网络结构及模板设计

神经网络主要分为三层：第一层将句子的每个字或类别关键词映射到字或类别向量空间。第二层是特征提取层。包括①通过滑动窗口机制选取字向量输入到神经网络提取字特征，即局部特征。②将句子的所有字向量输入到卷积神经网络层，最大池化后提取句子特征，即全局特征。③将句子中的类别关键词的类别关键词向量首尾相接输入到普通神经网络中提取类别特征。第三层将三类特征首尾相接作为全连接分类层的输入，最后预测字的标注。网络结构除重新构建外，可以通过加载预训练完成的网络作为初始模板。

（3）设置超参数并训练

超参数的取值对实验的结果有重要的影响，目前超参数的取值主要凭借实验、经验和随机初始化。训练终止方式有两种：一是设置训练代数，完成目标代数自动停止；二是设置性能阈值，达到某一性能阈值后自动停止。

（4）模型验证与性能评估

模型验证推荐采用以下三种方法：

①N折交叉验证：N折交叉验证也称循环估计，是一种统计学上将数据样本切割成较小子集的实用方法。将训练集分份，轮流将其中份作为训练数据，1份作为测试数据，进行试验。每次试验都会得出相应的正确率。次结果的正确率的平均值作为对算法精度的估计。

②留一法：每次从个数为的样本集中取出1个样本作为验证集，剩下的个样本作为训练集，重复进行次，依次取遍所有个数据作为验证集，最后将平均的个数据的结果作为泛化误差的估计。一般在数据缺乏时使用。

③留出法：留出法直接将数据集划分为两个互斥的集合，其中一个集合作为训练集，另一个作为测试集，即。在上训练出模型后，用来评估其测试误差，作为对泛化误差的估计。上述三种方法的特点和适用范围如表3-8所示。

表3-8 模型验证评估方法对比

|  |  |  |
| --- | --- | --- |
| 验证方法 | 特点 | 适用范围 |
| N折交叉验证 | Ｋ的取值会影响到最后评估结果的稳定性 | 数据量充足 |
| 留一法 | 在三种方法中评估更准确 | 数据集小且可有效划分 |
| 留出法 | 使用简单 | 数据量充足 |

模型性能评估常用根据召回率和准确率两个指标。若性能不满足要求则返回步骤3，调整参数后训练或增加训练代数。

（5）部署应用

调优后对模型进行简单的测试，最后将训练好的模型存储，并部署以供使用。



图2-7 基于人工智能的录入流程

3.3存储规范

知识的规则化输出结果主要是以图和表的形式，图知识和表知识有着不同的特性及结构，因此需要对它们分别设置不同的存储方案。以下对这两种形式知识的存储进行说明。

3.3.1 图知识的存储

图知识中有以下几个基本概念：图（Graph）指关系图。比如：同学及朋友关系图、银行转账图等；顶点（Vertex）一般指实体。比如：人、账户等；边（Edge）一般指关系。比如：朋友关系、转账动作等；属性（Property）顶点或边可以包含属性，比如：人的姓名、人的年龄、转账的时间。其逻辑形状如下图3-10所示。



图3-10 图知识逻辑形状

在图知识中最重要的信息就是图的节点、边的信息和边之间的关系，可以使用关系型数据库或图数据库来存储这些信息，具体如下。

图数据库是以图论为基础的一种非关系型数据库，其数据库存储结构和数据的查询方式都是以图论为基础的。图论中图的基本元素为节点和边，在图数据库中对应的就是节点和关系。图数据库的关注点是“关联关系”形成的图，其目标是对现实世界中的实体与实体之间的关联关系进行存储与分析：将实体抽象为顶点、将实体之间的关联关系抽象为边。通过顶点和边形成的图谱结构，直观自然地表达万物关联的世界，同时解决了复杂关联关系深层检索的性能问题。

图数据库仅使用一个模式和一组起点，就可以围绕这些初始起点探索相邻数据，收集和汇总来自数百万个节点和关系的信息，并保持搜索范围之外的任何数据不变。图数据库在处理规模大、关联度高的数据时优势明显。与传统的关系型数据库相比，图数据库具有如下优势：

（1）可以很自然地表达现实世界中的实体及其关联关系（对应图的顶点及边）；

（2）灵活的数据模型可以适应不断变化的业务需求；

（3）灵活的图查询语言，轻松实现复杂关系网络的分析；

（4）关系型数据库在遍历关系网络并抽取信息的能力非常弱，图数据库则为此而生；

（5）关系型数据库在规模庞大时很难做多层关联关系分析（Join操作往往消耗过长时间而失败），图数据库则天然把关联数据连接在一起，无需耗时耗内存的Join操作，可以保持常数级时间复杂度；

（6）在关系型数据库的查询中，关系的增多意味着需要遍历更多的关联表与实体表，其中存在着大量需要遍历的数据。并且，关系型数据库在数据查询时遍历的大量内容都与查询内容毫无关系，因此这也会导致关系查询时效率的大大降低；

（7）在进行数据查询时，图数据库会从查询目标的临近节点开始，基于图结构进行查询，而不是对整个类别的数据进行遍历。因此，在查询内容不变的情况下，数据规模的不断增大并不会对其查询性能产生明显影响；

（8）图数据库多用于网络安全、金融风控、知识图谱、广告推荐、社交网络、智能机器人、物联网等领域以及关联分析、路径搜索、子图挖掘、特征发现、社区聚类等应用。其主流的查询语言包括：Gremlin、Cypher、SPARQL，典型的存储结构有边集数组、邻接矩阵、邻接表和十字链表等。

综上所述，使用图数据库存储图数据能够产生更高的效率及性能。在使用图数据库存储图数据时，可以使用当前流行的图数据库Neo4j。Neo4j是一个由Java实现的开源图数据库，有效地将属性图模型实现到了存储级别，还提供了完整的数据库特性。它包括如下几个显著特点：完整的ACID支持、高可用性、轻易扩展到上亿级别的节点和关系、通过遍历工具高速检索数据。Neo4j确保了在一个事务里面的多个操作同时发生，保证数据的一致性。Neo4j把其数据库文件分为四大类来分类存储，分别是标签、节点、属性和关系。除此之外，Neo4j还提供了一个用户友好的Web页面，可以进行各项配置、写入、查询等操作，并且提供了可视化功能。Cypher是Neo4j的声明式图形查询语言，允许用户不必编写图形结构的遍历代码，就可以对图形数据进行高效的查询。Cypher的设计目的类似SQL，适合于开发者以及在数据库上做点对点模式（ad-hoc）查询的专业操作人员。其具备的能力包括：创建、更新、删除节点和关系；通过模式匹配来查询和修改节点和关系；管理索引和约束等。以上种种原因使得Neo4j最适合用于完整的企业部署或者用于一个轻量级项目中完整服务器的一个子集存在。

使用Neo4j图数据库存储图知识的步骤也非常简单方便：可以直接使用Cypher load csv语句将数据转换成CSV格式后通过LOAD CSV读取数据，也可以使用官方提供的neo4j-import工具进行批量导入存储。除了这两种方式之外，还可以使用官方提供的API或是一些别的导入工具进行存储。

3.3.2 表知识的存储

表知识中的各种数据被划分为一个个字段，每条数据对于同一个字段可能会有着不同的值，在存储表知识时可以使用关系型数据库。

关系型数据库（Relational Database）是建立在关系模型基础上的数据库，借助于集合代数等数学概念和方法来处理数据库中的数据。现实世界中的各种实体以及实体之间的各种联系均用关系模型来表示。关系型数据库以行和列的形式存储数据，用户通过查询来检索数据库中的数据，而查询是一个用于限定数据库中某些区域的执行代码。关系模型可以简单理解为二维表格模型，而一个关系型数据库就是由二维表及其之间的关系组成的一个数据组织。

关系型数据库具有如下几个特点：

存储方式：传统的关系型数据库采用表格的储存方式，数据以行和列的方式进行存储，要读取和查询都十分方便。

存储结构：关系型数据库按照结构化的方法存储数据，每个数据表都必须对各个字段定义好（也就是先定义好表的结构），再根据表的结构存入数据，这样做的好处就是由于数据的形式和内容在存入数据之前就已经定义好了，所以整个数据表的可靠性和稳定性都比较高，但带来的问题就是一旦存入数据后，如果需要修改数据表的结构就会十分困难。

存储规范：关系型数据库为了避免重复、规范化数据以及充分利用好存储空间，把数据按照最小关系表的形式进行存储，这样数据的管理就可以变得很清晰、一目了然。

扩展方式：由于关系型数据库将数据存储在数据表中，数据操作的瓶颈出现在多张数据表的操作中，而且数据表越多这个问题越严重，如果要缓解这个问题，只能提高处理能力，也就是选择速度更快性能更高的计算机，这样的方法虽然可以一定的拓展空间，但这样的拓展空间一定有非常有限的，也就是关系型数据库只具备纵向扩展能力。

查询方式：关系型数据库采用结构化查询语言（即SQL）来对数据库进行查询，SQL早已获得了各个数据库厂商的支持，成为数据库行业的标准，它能够支持数据库的CRUD（增加，查询，更新，删除）操作，具有非常强大的功能，SQL可以采用类似索引的方法来加快查询操作。

规范化：在数据库的设计开发过程中开发人员通常会面对同时需要对一个或者多个数据实体（包括数组、列表和嵌套数据）进行操作，这样在关系型数据库中，一个数据实体一般首先要分割成多个部分，然后再对分割的部分进行规范化，规范化以后再分别存入到多张关系型数据表中，这是一个复杂的过程。随着软件技术的发展，相当多的软件开发平台都提供一些简单的解决方法，例如，可以利用ORM层（也就是对象关系映射）来将数据库中对象模型映射到基于SQL的关系型数据库中去以及进行不同类型系统的数据之间的转换。

事务性：关系型数据库强调ACID规则（原子性（Atomicity）、一致性（Consistency）、隔离性（Isolation）、持久性（Durability）），可以满足对事务性要求较高或者需要进行复杂数据查询的数据操作，而且可以充分满足数据库操作的高性能和操作稳定性的要求。并且关系型数据库十分强调数据的强一致性，对于事务的操作有很好的支持。关系型数据库可以控制事务原子性细粒度，并且一旦操作有误或者有需要，可以马上回滚事务。

读写性能：关系型数据库十分强调数据的一致性，并为此降低读写性能付出了巨大的代价，虽然关系型数据库存储数据和处理数据的可靠性很不错，但一旦面对海量数据的处理的时候效率就会变得很差，特别是遇到高并发读写的时候性能就会下降的非常厉害。

在关系型数据库中，数据由不同的表组成，每个表包含有不同的字段，还可以对各个字段施加多种限制，其结构如下表3-9所示。

表3-9 关系型数据表结构

|  |  |  |
| --- | --- | --- |
| 字段1 | 字段2 | 字段3 |
| A | E | I |
| B | F | J |
| C | G | K |

在使用关系型数据库存储图数据时，可以使用当前流行的MySQL数据库管理系统。在Web应用方面，MySQL是最好的RDBMS （Relational Database Management System，关系数据库管理系统） 应用软件之一。MySQL所使用的SQL语言是用于访问数据库的最常用标准化语言。MySQL软件采用了双授权政策，分为社区版和商业版，由于其体积小、速度快、总体拥有成本低，尤其是开放源码这一特点，一般中小型网站的开发都选择MySQL作为网站数据库。

使用MySQL数据库存储表知识也是非常方便快速：提前对表知识进行一些规范化处理后将其存储为CSV文件、EXCEL文件等格式，然后可以使用各种各样的数据库管理工具从这些文件中批量导入数据即可。

3.3.3 存储工具推荐

综上所述，图知识和表知识有着不同的特性和结构，其存储方式、存储工具等都不相同，对它们的存储既要符合它们独特的数据结构，又要能够在查询时方便地进行检索。关于图知识和表知识的特点、可使用的存储工具及建议使用的较成熟的数据库工具如下表3-10所示：

表3-10 图知识和表知识特点及建议存储工具

|  |  |  |  |
| --- | --- | --- | --- |
| 数据类型 | 特点 | 可使用的存储工具 | 建议使用 |
| 图知识 | 数据以节点和边的形式展现出来 | 图数据库、关系型数据库 | Neo4j图数据库 |
| 表知识 | 数据以二维表格的形式展现出来 | 关系型数据库 | MySQL数据库 |

3.4知识规则化示例

本部分针对风洞基地的一个故障案例——“水泵运转不上水”，基于知识图谱将其中涉及的知识规则化，对知识之间的联系进行分析和描绘。案例主要描述如下：

故障：水泵运转不上水。

时间：2019年6月14日；故障发现人：杨某。

故障/事故描述：试验运行前，打开为热阀供水水泵，水泵运转起来，水泵出口压力不上升。

检查经过：拆解水泵底阀，发现底阀封面有沙粒，底阀密封不严。

发生原因：水泵长期使用，水质变差，水中杂质寄存于底阀密封面，造成底阀无法密封，水泵抽吸为抽空状态，导致无法上水。

依据知识图谱的应用流程，从上述案例中进行知识抽取，得到相关实体、关系、属性以及属性值，即三元组（实体，关系，实体）或（实体，属性，属性值）的元素，如表3-11所示。用箭头将所有三元组连接起来，即可得到如图3-11所示的水泵故障知识图谱。

表3-11 知识抽取结果

|  |  |
| --- | --- |
| 类型 | 结果 |
| 实体 | 水泵、杨某等 |
| 关系 | 故障结果、措施等 |
| 属性 | 发现时间 |
| 属性值 | 2019年6月14日 |



图3-11 水泵故障知识图谱

在得到上图中的知识图谱后，可以使用图数据库Neo4j将其存储起来。Neo4j使用Cypher查询语言对数据库中图数据进行处理，能够方便地构建、查询图知识。上图中的知识图谱可使用下列Cypher语句进行构建：

create

(:杨某{name:"杨某"}),

(:运转不上水{name:"运转不上水"}),

(运转不上水)-[:发现人]->(杨某),

(:更换水体{name:"更换水体"}),

(运转不上水)-[:措施]->(更换水体),

(:清除阀门杂质{name:"清除阀门杂质"}),

(运转不上水)-[:措施]->(清除阀门杂质),

(:时间{name:"2019年6月14日"}),

(运转不上水)-[:发现时间]->(时间),

(:水泵{name:"水泵"}),

(水泵)-[:故障结果]->(运转不上水),

(:出口压力不升{name:"出口压力不升"}),

(水泵)-[:故障描述间]->(出口压力不升),

(:拆解底阀{name:"拆解底阀"}),

(水泵)-[:检查方式]->(拆解底阀),

(:底阀无法密封{name:"底阀无法密封"}),

(出口压力不升)-[:原因]->(底阀无法密封),

(:水中杂质{name:"水中杂质"}),

(底阀无法密封)-[:原因]->(水中杂质)

在Neo4j数据库中使用以上Cypher语句构建之后产生的图知识逻辑结构如下图3-12所示，之后如果需要对知识图谱中内容进行查询或修改即可使用Cypher语句快速完成。

图表

描述已自动生成

图3-12 在Neo4j数据库中的水泵故障知识图谱

4知识应用规范

规范应用知识是创造专家知识价值的前提。上一章中阐述了知识录入与存储规范，本章是知识应用规范。首先，规范了知识库设计的基本内容，随后，提出了知识应用的评价准则，对平台中的知识使用率、点赞数等进行统计，体现出知识的使用情况，进而通过不断的迭代，提升知识使用效率。最后，阐述了知识维护的必要的规范流程及内容，指导知识库后续的维护。

4.1知识库设计

知识库是知识、数据库技术相结合形成的系统，构建知识库是故障诊断等应用的基础，是推理决策的支撑。首先需要对源知识进行管理，以便于源知识复用、索引、检查。从源知识抽取并规则化后形成规则和事实两部分，需要规范两者的编码。

4.1.1源知识数据管理

（1）通用要求

①具有唯一性；

②采取分段式编码。

③采用字母缩写的编码，如针对图片的格式的存储可以被定义为FIG。

（2）字符要求

①数字字符“0”至“9”；

②大写英文字母“A”至“Z”，但不包括“I”和“O”（易与数字“1”和“0”混淆）。

（3）编码设计

依据Q/QD 9-2021对气动中心装备管理编码规则的定义，规则名编码由“管理机构+所属系统编码+文件类型 +时间序号”组成。分段编码规范如图4-1所示。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 510×× | - | ×××-××× | -××× | - | ××× |
| 管理机构码（前5位） |  |  |  |  |  |  |
| 所属系统类别及序号 |  |  |  |  |  |  |
| 文件类型 |  |  |  |  |  |  |
| 时间序号 |  |  |  |  |  |  |

图4-1 源知识/材料编码规范

管理机构码：表示管理结构的识别码如低速空间动力研究所的标识码51013。

所属系统序号：表示部门管理的具体风洞类别及序号，以“字母”+“数字”组成,如4米×3米低速风洞设备可以被表示为“AFL12Z-0104001”。

文件类型：表示源知识/材料中的图、表等文件类型，如图片可以被表示为“FIG”，表可以被表示为“TAB”，声音可以被示为“VIC”，word文档可以被表示为“DOC”。

时间序号：以时间序号的方式来表示该文件的时间戳及相应的序号，如2021年9月23日第一个文件可以被表示为2021092300001。

举例：如低空气动力研究所4米×3米低速风洞风扇所产生的2021年9月16日第一个word文档可以被表示为51013-AFL12Z-0104001-DOC2021091600001。

4.1.2规则名和事实名的编码规范

（1）通用要求

①具有唯一性；

②采取分段式编码；

③采用字母缩写的编码，如针对电机的规则可以被简写成DJ；

④为了区分规则名和事实名，在分段式的编码中，采用“GZ”和“SS”来进行区分。

（2）字符要求

①数字字符“0”至“9”；

②大写英文字母“A”至“Z”，但不包括“I”和“O”（易与数字“1”和“0”混淆）。

（3）编码设计

依据Q/QD 9-2021对气动中心装备管理编码规则的定义，规则名编码由“管理机构+所属系统编码+规则名/事实名+设备序号+规则/事实号”组成。分段编码规范如图4-2所示。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 510×× | - | ×××-××× | - | ××- | -××× |
| 管理机构码（前5位） |  |  |  |  |  |  |
| 所属系统类别及序号 |  |  |  |  |  |  |
| 规则名/事实名 |  |  |  |  |  |  |
| 规则/事实号 |  |  |  |  |  |  |

图4-2 规则/事实编码规范

管理机构码：表示管理结构的识别码如低速空间动力研究所的标识码51013。

所属系统序号：表示部门管理的具体风洞类别及序号，以“字母”+“数字”组成,如4米×3米低速风洞设备可以被表示为“AFL12Z-0104001”。

规则名/事实名：区分规则和事实，用大写字母缩写“GZ”和“SS”区分。

规则/事实号：具体的规则或者事实序号，如温度高导致电机停止运转的编码可以用“WDH”和“ST”表示。

举例：低速空气动力研究所管理的4米×3米低速风洞风扇产生的温度高故障导致电机停止运行的规则名或者事实名可以分别被表示为51013- AFL12Z-0104001-GZ-WDH，51013- AFL12Z-0104001-SS-ST。

4.2知识应用评价准则

知识应用评价准则用于评价录入的知识，反映录入知识的使用、更迭情况，并规范和修正录入的知识，如统计性准则。

统计性准则主要是通过对录入知识增加的使用次数、点赞数、反馈、版本号的属性进行统计和更新，来评价录入的知识的价值。当该录入知识被检索使用时，“使用次数”递增；当录入知识被修改更迭时，版本号被更新；当该录入知识被使用的时候，可以增加对该知识的使用反馈。首先，为了能够简单高效地使用上述的统计性属性，利用“关系图”数据结构去建立一条录入知识和使用次数、知识使用反馈、版本号等之间的连接，如下图4-3所示。

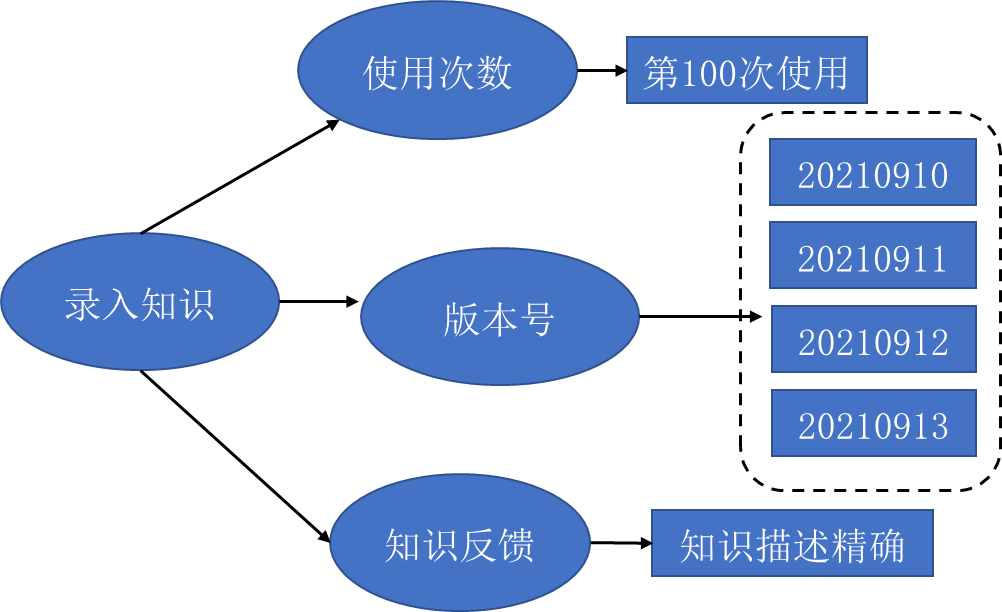


图4-3 录入知识的统计属性（部分）

其次，根据录入知识的“关系图”，建立相应的统计准则表4-1，通过访问录入知识的属性可以判定该录入知识的整体使用、更新情况。

表4-1 知识统计属性表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 知识ID | 录入知识 | 知识反馈 | 使用次数 | 错误次数 | 点赞数 | 版本号 |
| 1 | 若XXX，则A | 描述精确 | 10 | 0 | 10 | 20210910 |
| 2 | 若XXX，则B | 特殊情况xxx不支持，需要补充 | 1 | 1 | 0 | 20210911 |
| 3 | 若XXX，则C | 描述较精确 | 10 | 9 | 0 | 20210912 |
| 4 | 若XXX，则D | 特殊情况xxx不支持，需要补充 | 3 | 2 | 1 | 20210913 |
| 5 | 若XXX，则E | 描述错误 | 1 | 1 | 0 | 20210914 |

最后，通过访问某录入知识可以访问到该表，并计算出相应的统计指标。例如，通过计算某录入知识错误的次数占使用次数的比例（知识ID为1的错误比例是0）来评价录入知识的描述是否精确。知识应用评价数据库字段设计如表4-2所示。

表4-2 数据库知识应用评价表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据库字段名 | 数据库字段编码 | 数据库字段类型 | 数据库字段长度 | 是否为主键 | 是否可以为空 | 是否自增 |
| 知识ID | UTF-8 | NUMBER | 10 | 是 | 否 | 是 |
| 录入知识 | UTF-8 | VARCHAR | 255 | 否 | 否 | 否 |
| 知识反馈 | UTF-8 | VARCHAR | 255 | 否 | 否 | 否 |
| 使用次数 | UTF-8 | NUMBER | 10 | 否 | 否 | 否 |
| 错误次数 | UTF-8 | NUMBER | 10 | 否 | 否 | 否 |
| 点赞数 | UTF-8 | NUMBER | 10 | 否 | 是 | 否 |

4.3知识维护流程

由于知识不是一成不变的，知识的应用领域随时间发生变化，同时有关术语的说明也在发生变化。因此，对于已经构建的知识库需要进行维护，当知识发生变化的时候系统应提供及时修改的工具。此外，新加入的知识是否与知识库中的原有知识能保持一致性，是否存在潜在的冲突也是一个非常关键的问题。所以知识维护成为知识库管理的重要内容，其目的是保证知识的一致性、完备性、低冗余性。知识的一致性、完备性、低冗余性三者并非完全割裂，而是相辅相成，共同作用于知识维护整个流程。而知识库通过各种途径获得的知识，必须经过检查和调整才能和原有知识成为一个整体，从而保证知识库的可用性。

知识维护的另外一个非常重要的工作是进行相容性检查，即检查新加入的知识是否与已有的知识矛盾，以保证整个知识集合的一致性或称相容性。对于新老知识不一致有两种处理方法：一是以新加入的知识为准来调整即适当地修改己有的知识，使之成为一致的方法称为调整；二是适应以已有的知识为准来适当修改新知识，称为同化。

4.3.1一致性检测

一致性检测是通过语义分析的方法将待检测标准与已有标准知识库进行比较，从而发现不一致。一致性检测可以描述为三个步骤：建立标准知识库、待检测标准抽取、一致性检测，如图4-4。



图4-4 一致性检测步骤

标准知识库是标准一致性检测的基础，是一系列标准内容的数据表现。在建立标准知识库时需要从标准体系中选择标准作为基准，从这些基准标准中抽取出来信息，以一定格式存储在数据库中，为后续的检测工作做准备。每一个基准标准的抽取方法与待检测标准的抽取方法相同。

要对标准进行检测，首先要把标准的各部分准确地抽取出来，这也是整个研究的重点和难点之一。抽取包括两部分含义：其一是对标准文档进行内容抽取，得到标准各部分；其二是对部分标准内容进行信息抽取，以适应后续的检测步骤。

从标准中抽取出来条款以后，就要对其进行一致性检测。一致性检测是将待检测标准的条款与标准知识库中保存的各种基准标准的条款进行相似度比较、核准。发现不一致的地方，根据标准定义的内容，进行判别和统计。

知识的一致性包括语法一致性和语义一致性两个方面。语法一致性是指知识

单元记录格式的标准化和规范化。语义一致性是指知识间不存在冗余、矛盾，使之能经过有限的推理步骤而获得正确的结论。其检验方法是在新规则输入之时，将新规则与所属子知识库中的原有规则逐条比较来进行。一般来说，每当有新规则纳入知识库时，就应该对新规则和原知识库中旧规则之间可能存在的不一致性问题进行检测。如果这种检测等到知识库膨胀到相当规模时才进行，那么，由于问题的积累而形成的错综复杂的矛盾局面，会使这种检测和相应的改进无法进行。其算法流程如图4-5所示。

（1）冗余

当某一规则可用另外的规则表达或代替时，称该规则是冗余的。包括以下几种情形(式中“^”表示“与”):

①等价规则：一条规则的前件和后件与另一条规则的前件和后件完全等价。

例如： 

这时，R1、R2是等价规则，应删除其中一条规则；同样，R3、R4是等价规则，应删除其中一条规则。

②从属规则

从属规则可分为两种情况：



图4-5 冗余性验证算法

a.条件从属：两条规则R1和R2前件相同，但R1的前件部分约束比R2多，则R1是R2的从属规则。

例如： 

在这种情况下，要依据处理问题的策略来确定在何时用较一般的规则，何时用较特定的规则。

b.结论从属：两条规则R1和R2前件相同，但R1的后件部分约束比R2多，则R1是R2的从属规则。

例如： 

在这种情况下，一般删除规则R2。

c.传递冗余：若两条规则链中第一条规则前件相同，而最后一条规则后件是等价的，则称这两条规则链是传递冗余的。

例如： 

当Q(x)是推理过程中用到的中间结果或解释过程中用到时，保留前者，否则保留后者。

（2）矛盾

知识库中存在矛盾的规则或规则链，则称知识库中含有矛盾。两条规则或规则链，当且仅当它们在相同的前件下得到后件中有相互矛盾的结论时，称这两条规则或规则链是矛盾的。主要有以下几种情况(式中“~”表示“非”)：

①自相矛盾：从某一前提直接推得与之相反的结论，成为自相矛盾。

例如：

②传递自相矛盾：从某一前提P(x)经过一推理链推出相反的结论~P(x)。

例如：

③隐含矛盾：隐含矛盾有两种情形。

a.在知识库中存在一推理链，从推理链后面某一节点中能推出与前面节点相反的结论，但这一规则未包含在知识库中。

例如：，且规则应当包含在知识库中，但未显式地包含在知识库中。

b. ，，且应当包含在知识库中，但是漏掉了。

④相互矛盾：两条规则有相同的前件，但后件存在相互矛盾的结论。

例如： 

⑤传递矛盾：若两条规则链中的一条规则前件相同，而最后一条规则的后件中存在相反的结论，则称知识库中存在传递矛盾。

例如： 

（3）循环

知识库中，如果存在一规则R1，R1的后件中含有与R1的某前提相同的结论,或存在一规则链“R1，R2，R3…Rn”，Rn的后件中含有与R1的某前提相同的结论，那么，则称知识库中存在循环。主要有以下几种情形：

①在某规则的前件和后件中引入同一谓词。

例如：

②规则链中出现自我引用。

例如：

③隐含循环

例如：且应当包含在知识库中，却未显式地包含在知识库中。

4.3.2完备性维护

如果知识管理系统能处理在运行过程中遇到的所有问题，那么称该知识管理系统的知识库是完备的。但是，在知识管理系统开发的初期，这是很难做到的。例如，获取的知识不全面、知识表达不全面，以及编码过程中的遗漏和错误都可导致知识库的不完备。主要有以下几种情况：

（1）多余输入：系统向用户询问P(A)，但P(A)未在任何规则的前提中出现。

（2）死节点：在知识库中存在规则，且Q(x)不是一个目标，且不在前集中出现。

（3）不可达目标：目标R(x)不在结论集中，或者在知识库中存在规则，且Q(x)是不可达的(这一过程是递归的)。

（4）未用的属性值：值全部或部分地未被任何规则引用。

（5）遗漏规则：应该存在于知识库中，但尚未包含在知识库中的规则，称为遗漏规则。

4.3.3知识更新

知识更新的一般方法是进行知识库的实体扩充。实体扩充的目标是从网络大数据的文本中获取的实体动态扩展到知识库中，从文本中获取的实体与知识库中的实体存在两种关系：一是知识库中存在与文本实体映射的实体，对此类实体只需要找到文本实体在知识库中的映射实体，即实体链接。二是知识库中不存在与文本实体映射的实体，首先基于知识库中分类为文本实体标注类别，即实体分类，然后根据分类将文本实体护展到知识库对应的分类下，完成文本实体与知识库的关联合并。实体扩充方法如图4-6所示。



图4-6 实体扩充方法

（1）实体链接方法

实体链接的主要作用是利用知识库的实体对从文本中获取的实体指代进行消歧，识别每个实体指代在知识库中与其对应的映射实体。按照实体链接采用的信息不同，现有的实体链接方法具体分为以下四种。

①基于实体属性的实体链接方法。早期的实体链接方法通过计算实体的属性相似度判断实体是否相同，最直接的方法是基于实体名字属性的字符串相似度方法，但是基于字符串相似度的方法无法处理实体语义异构的情况，如“凤梨”和“菠萝”表示同一实体，基于字符串相似度的方法会将其判别为两个不同实体。在此基础上，通过词干还原，查找该词干在语义词典中的同义词集合表示和描述，通过词语在语义词典中的概念层次结构中的最短路径、同义词集合和描述计算语义相似度，将实体的字符串相似度和语义相似度进行加权。

②基于实体流行度的实体链接方法。基于实体流行度的方法本质上是一种基于概率统计的方法，所基于的假设条件是对于给定的实体指代，与其对应的映射实体大概率出现现实中最流行的实体。例如，给定一个实体指代“李娜”，人们可能最先想到的是著名网球运动员“李娜”。

③基于实体上下文的实体链接方法。这一方法通过计算与实体相关的上下文的相似性来判断两个实体是否同实体。这一方法基于的假设条件是如果两个实体的上下文相似，那么这两个实体就可能是同一个实体。

基于实体上下文的方法可以弥补基于实体流行度方法的缺陷，但是上下文相似度方法要求两个被比较的文本同存在词重叠，由于自然语言使用的灵活性，这会成为个较为严格的来条件。

④基于外部证据的实体链接方法。这一方法通过采用话题连贯性提升实体链接的准确性，将候选实体和同一上下文中其他实体的分类和链接的重叠率计算实体间的话题连贯性，该方法认为同一文本的实体并不是独立的，存在着语义相关性，这种相关性有助于提升实体链接的准确性。

（2）实体分类方法

实体分类的主要目标是对大数据的文本中获取的实体进行类别标注，按照分类标注粒度不同，实体分类方法有粗粒度方法和细粒度方法。

①粗粒度的实体分类方法。粗粒度的实体分类主要是将实体分为人名、地名、机构名等类别。在粗粒度的分类中占主导地位的是有监督的方法，如著名的汉语词法分析系统采用层叠隐马尔可夫模型将汉语词法分析的所有环节都统一到个完整的理论框架。半监督方法的主要思想是利用种子训练数据，通过自我学习不断标注新的样本数据，迭代改进分类方法的准确性。无监督的方法主要思想是实体分类在没有任何标注数据的条件下进行，计算给定的实体上下文与主题签名之间的相似度，利用相似度最高的签名对实体进行分类标注。

②细粒度的实体分类方法。细粒度的实体分类则根据本地或知识库包含的成干上万分类信息对实体进行更细致的类别标注.传统的细粒度实体分类方法是有监督的基于分类模型的方法，通过提取一些语言特征，如词、词性和实体上下文等分类训练器。半监督的方法更适合于细粒度的实体分类，利用本体中与实体上下文相似度最高的分类标注实体。无监督的方法通过实体上下文建立文本实体与知识库之间的关系，在创建的关系图上利用基于随机游走的标签传播算法获得文本实体的分类。将两类实体分类的方法汇总到表4-3。

表4-3 实体扩充中的实体分类方法汇总表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 方法 | 主要特点 | 学习方式 | 优点 | 不足 |
| 粗粒度的实体分类方法 | 实体类别标签粒度相对较粗，主要将实体分为人名、地名、机构名等大类别 | 有监督 | 可充分利用分类分布的先验知识，控制训练样本的选择。获得较高的准确率 | 人为主观因素较强，训练样本的选取和评估需要花费较多人力与时间 |
| 半监督 | 能自动在无标注的样例的帮助下训练有类别标签的样本，弥补调练样本不足的缺陷 | 采用的样本数据都无噪声干扰。但在实际中难以得到纯样本数据，导致分类准确率降低 |
| 无监督 | 不需要人工标注训练数据，人为误差的机会减少，需输入的初始参数较少 | 无需对实体类别划分有较多了解，输入的初始参数较少，所分的类别比监督的方法更均质 |
| 细粒度的实体分类方法 | 实体类别标签粒度相对较细，将实体可能分为成千上万个实体类别 | 有监督 | 控制训练样本的选择，并可通过反复  检验训练样本，提高分类的准确率 | 标注具有成干上万类别的训练样本需要花费大量的人力与时间 |
| 半监督 | 自动对未标记数据加以利用，学习整个数据分布上具有较强泛化能力的模型 | 要求参与分类的实体没歧义或实体在语料中都表示为统一规范的形式，这在实际应用中难以得到满足 |
| 无监督 | 无需对实体类别划分有较多了解，输入的初始参数较少，所分的类别比监督的方法更均质 | 实体的分类结果仍需大量的分析和后处理，与有监督学习相比，实体分类的时间复杂度偏高 |

4.3.4知识备份

对于知识库的维护还包括对知识库的备份和还原工作。在某些情况下，需要对系统中的知识库进行备份。备份的目的在于便于知识库的保存和传递。数据库的备份将主要使用到数据的相关功能，在代码中可直接调用数据库备份模块。首先由用户指定备份文件的位置和名称，然后调用备份模块，生成最终备份文件。和数据库备份相反的操作是对知识库的还原，通过还原操作，可将已有知识库文件重新导入到数据库系统中。

常用的知识备份方式有三种，如表4-4所示：

（1）完全备份

备份全部选中的文件夹，并不依赖文件的存档属性来确定备份哪些文件。

在备份过程中，任何现有的标记都被清除，每个文件都被标记为已备份。换言之，清除存档属性。

这种备份方式最大的好处是只要用一个备份，就可以恢复数据。因此大大加快了系统或数据的恢复时间。

（2）差异备份

备份自上一次完全备份之后有变化的数据。

差异备份过程中，只备份有标记的那些选中的文件和文件夹。它不清除标记，即备份后不标记为已备份文件。换言之，不清除存档属性。

差异备份是指在一次全备份后到进行差异备份的这段时间内，对那些增加或者修改文件的备份。在进行恢复时，我们只需对第一次全备份和最后一次差异备份进行恢复。差异备份在避免了另外两种备份策略缺陷的同时，又具备了它们各自的优点。

首先，它具有了增量备份时间短、节省磁盘空间的优势；其次，它又具有了全备份恢复所需磁带少、恢复时间短的特点。

（3）增量备份

备份自上一次备份（包含完全备份、增量备份）之后有变化的数据。

增量备份过程中，只备份有标记的选中的文件和文件夹，它清除标记，即：备份后标记文件，换言之，清除存档属性。

增量备份是指在一次全备份或上一次增量备份后，以后每次的备份只需备份与前一次相比增加和者被修改的文件。这就意味着，第一次增量备份的对象是进行全备份后所产生的增加和修改的文件；第二次增量备份的对象是进行第一次增量备份后所产生的增加和修改的文件，如此类推。

这种备份方式最显著的优点就是：没有重复的备份数据，因此备份的数据量不大，备份所需的时间很短。表4-4对上述各类备份方式进行了比较。

表4-4 备份方式比较

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 知识备份方式 | 备份内容 | 数据量 | 是否标记 | 备份所需时间 | 恢复所需时间 |
| 完全备份 | 所有内容 | 数据量大（全部数据） | 全部标记 | 长 | 短 |
| 差异备份 | 上次**完全备份之后**有变化的内容 | 少于完全备份，多于增量备份  （多个差异备份之间有重复数据） | **不标记** | 中 | 中 |
| 增量备份 | **上次备份**后有变化的内容 | 数据量小（没有重复的备份数据） | 标记 | 短 | 长 |

4.4知识维护示例

冗余数据检测算法分为3个步骤：数据分组、字段匹配、冗余判断。数据分组就是对数据进行聚类操作，将可能相似重复的数据集中至一定区域，随后在给定区域内进行数据匹配，极大程度的缩减搜索空间。其中，最具代表性的是Erhard Rahm等人提出的近邻排序算法（Sorted-Neighborhood Method，SNM）。后续出现的多趟近邻排序算法、优先队列算法也是利用这一思想。字段匹配是指在特定区域内通过字段匹配算法计算记录之间的相似度分数。冗余判断指根据相似度分数和冗余阈值判断记录是否冗余。

对海量数据集进行划分，形成小数据集；其次对每个小数据集采用等级综合评价法为属性设置权重，以权重大小排序关键字；在字段匹配阶段，结合Q-Gram静态索引修建技术和Levenshtein编辑距离算法进行相似判断；最后采用可伸缩的滑动窗口进行验证。

（1）数据分组

数据分组中的近邻排序算法的伪代码如表4-5所示。

表4-5 近邻排序算法伪代码

|  |
| --- |
| 1：*Input*：数据集D，排序关键字*k*[]，滑动窗口大小*Q*，相似度阈值*L*.  2：*Output*：相似重复记录对*dup*[].  3：构建*SNM*模型  4： *for xi in D do*  5： 根据排序关键字对数据集D进行排序  6： *for xi in Q do*  7： 将窗口*Q*中的第一条记录与窗口中其余记录依次比较  8： 将相似度分数大于给定阈值*L*的记录判定为冗余数据，进行输出  9：  *end for*  10： *end for*  11：*return dup* |

SNM算法利用滑动窗口精减了记录间的比较次数，使匹配效率得以提升。SNM算法的时间复杂度为，其中，为窗口大小，为数据中的记录总数。针对海量数据处理时间效率低下的问题，可将海量大数据集划分为若干个不相交的小数据集，划分方法如下。

步骤1：针对不同数据集的不同特性，以专家经验分析为前提，选取具有代表性的某个属性，以该属性把大数据集分割为若干个不相交的小数据集，称为簇。将大数据集记作D，属性,划分为*n*个不相交的子集，即。

步骤2：若某些划分后的数据集的数据量仍超过8000，则再次对该数据集实施划分。对数据集，选取属性,依据属性划分为*k*个不相交的子集，表示为。

步骤3：若仍存在某些数量大于8000数据集，可重复步骤2，对数据集进行迭代划分，直至各小数据集数量控制在8000以内。

数据划分依赖于问题的求解，为了求解的精确度，可将只选取单一属性进行数据划分，扩展到多属性选取进行数据划分。对仅有一条数据的簇，进行剪枝处理。

（2）字段匹配

由于关键字选取对相似重复记录检测效率和检测精度都有着重大影响，采用等级综合评价法结合客观的数理统计方法和主观的专家经验，综合考虑了各个属性对记录的贡献程度，使关键字的选取更为合理。

每个属性都有值域。多次在数据集中随机选取相同大小的样本数据，统计属性的长度，称为属性值。为减少随机取样时样本质量差异对属性值的影响，因此以均值法确定数据集中各属性值，若属性值越大，此属性的记录差异越大，该属性所占权重也就越大。属性值统计如表4-6所示。

表4-6 属性值统计表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 随机取样次数 | 属性P1 | 属性P2 | 属性P3 | … | 属性Pm |
| 1 | *Y11* | *Y12* | *Y13* | *…* | *Y1m* |
| 2 | *Y21* | *Y22* | *Y23* | *…* | *Y2m* |
| *…* | *…* | *…* | *…* | *…* | *…* |
| *n* | *Y31* | *Yn2* | *Yn3* | *…* | *Ynm* |
| 属性值 | *Y1* | *Y2* | *Y3* | *…* | *Ym* |

采用均值法确定属性*pj*的经验等级值*Gj*。



其中，表示第个用户为第个属性的分配的经验等级。

根据上述分析，属性的等级越高，其重要性越高，所对应的属性的权重也就越大，将数理统计方法与经验等级法相结合，计算得到最终的综合属性权值。



其中，，表示记录中有个属性。

等级综合评价法伪代码如表4-7所示

表4-7 等级综合评价法伪代码

|  |
| --- |
| 1：*Input*：用户数*n*，用户经验*G*，选取次数*time*，随机选取记录数*count*.  2：*Output：(W1,W2,…,Wm)*(*m*表示记录的*m*个属性).  3：*computeWeight(){*  4：*Y=statisticsWeight(time,count)*;//统计  5：*G=gradeWeight(n,E)*;//经验  6：*for (i =1…m){*//计算最终的权重  7：*Wi=1/2\*(Yi + Gi)*;  8：}  9：*return W*;  10：} |

依据上述等级综合评价法，操作数据集以属性对应的权重W从大到小进行排序，选出前k个属性作为数据集的排序关键字。

（3）冗余判断

索引技术通过在决策模型进行显式比较之前，删除明显不匹配记录来减少候选集的数量以实现效率，因此索引技术常被用于提高大型数据集匹配效率和可扩展性。

Q-Gram索引技术由于在构建倒排索引期间会生成大量冗余候选集，导致整体运行时速不高，但Q-Gram索引却具有可靠的检测准确率。因此采用静态索引裁剪技术在Q-Gram构建倒排索引后立即减小索引大小，在保证检测准确率的同时，减少查匹配处理时间。

每个q-gram标记作为索引键插入到反向索引中,直到当前窗口内所有记录都揷入到倒排索引后,修剪出现频率为1和频率超过阈值的q-gram标记,这两种情况分别表示该q-gram标记无匹配和该q-gram标记是公共块,因此进行修剪,避免过多不必要的匹配。一旦建立统一修建的倒排索引,选取当前窗口中第一条记录,根据阻塞秘钥值生成q-gram标记列表,若q-gram标记所包含的记录数大于阈值,则将该q-gram标记从的q-gram标记列表中移出,确保仅处理少量q-gram标记,以提高记录匹配效率。将与具有相同q-gram标记的记录放入,然后将中的所有记录与顺序比较。采用Levenshtein编辑距离算法计算属性间的编辑距离以及相似度,记录与记录在属性上的相似度计算公式



其中,表示记录与在属性上的编辑距离,表示记录中属性取值的字符串长度。若,则将该属性相似度分数置为1，反之,设置为0。最后，通过属性权重来缩放属性相似度分数,将所有加权属性相似度分数总和与阈值进行比较，鉴别相似重复记录对。一旦已检测完成的各属性相似度分数之和大于阈值，则停止检测剩余属性，减少不必要的属性间匹配。基于静态索引修剪技术的字段检测算法整体流程伪代码如表4-8所示。

表4-8 基于静态索引修剪技术的字段检测算法的伪代码实现

|  |
| --- |
| 1：*Input*：以*BKV*为中心的阈值*tu*,以记录为中心的阈值*td，*属性相似度阈值*fsim*,记录相似度阈值*rsim*,各属性权重*wf*.  2：*Output*：相似重复记录对.  3：*TermCentricStaticIndexPruning*(*double* *tu*)//倒排索引静态修剪  4：*LOOP*: *Inverted* *Index* *key*, *q*  5：*IF:|q| == 1 THEN* //无匹配(|*q*|表示该*q-gram*标记的频率)  6： *Delete q form Inverte dIndex*  7：*ELSEIF:|q|>(N\*tu)THEN//q-gram*标记包含倒排索引的记录过多  8： *Delete q from Inverted Index*  9：*END IF*  10：*END LOOP*  11：*end BKV CentricStaticIndexPruning*  12：*EfficientRecordMatching(double td, double fsim, double rsim, double wf)*//以记录为中心的阈值*td*,属性相似度阈值*fsim*,记录相似度阈值*rsim*,各属性权重*wf*  13：*LOOP: pick the first record rc from Window*  14： *Extract list q of q-gram tokens from BKV of rc*  15：  16： *IF:|q|>ld THEN*  17： *Remove extracted list of q-grams q*  18： *END IF*  19： *Extract list of unique records rx from q into Lr*//将与*rc*具有相同*q-gram*标记的*rx*放入*Lr*中  20： *LOOP: choose rx from Lr*  21： *LOOP: available attribute fr in both(rc,rx)*//循环比较记录*rc*与*rx*属性间的相似度  22：  23： *IF*: *THEN* 记录*rc*与 *rx*在属性*f*上的相似度大于属性相似度阈值，则该属性相似度得分置为1  24：  27： i  28：  //记录rc与rx的相似度得分为各属性相似度得分与其对应权重的乘积和  29： *IF*:  30： *Pair(rc,rx) as matched records*  31： *END LOOP*  32： *END IF*  33： *END LOOP*  34：*END LOOP*  35：*end EfficientRecordMatching* |

传统近邻排序算法滑动窗口大小难以设置，窗口过大或者过小都会出现一系列的问题，从而影响最终的检测效果，因此滑动窗口大小的设置也极为重要。本文根据窗口内记录间的相似程度动态地调整滑动窗口大小。记窗口最大值为Qmax，窗口最小值为Qmin，当前滑动窗口大小在Qmin和Qmax之间变化，滑动窗口大小根据记录相似度的计算值与阈值的比较进行灵活调整。可伸缩滑动窗口需设置3个参数，窗口最小值Qmin，窗口最大值Qmax，窗口最小阈值LowT，以及变量当前滑动窗口大小Qi。窗口初始值Qi设定为Qmin，窗口中的第1条记录R1首先在Qmin范围内与其他记录进行匹配，当匹配到记录RQmin时，如果相似度，说明此时窗口内记录间相似程度较高，应扩大窗口继续进行匹配，窗口扩大为：



如果相似度 ,记录对的相似度较低，应缩小窗口，减少不必要的比较。此时窗口*Qi*缩小为：



最终基于滑动窗口可实现冗余检测功能。

5初步应用建议

如何利用前述流程、技术，结合基地知识管理现状，对基地试验设备健康管理产生的知识实施规范化管理，是下一步建设知识库面临的重要问题。目前，需要进一步研究现存知识的特性，在此基础上，依据报告整体流程针对现存知识提出知识管理初步建议。

5.1现有知识类型分析

经过调研，基地现有知识主要包括规则知识、故障树和案例知识三种类型，其中，现存的规则知识为产生式规则化方法的结果，故无需进一步处理，可以继续使用；故障树具有充分的逻辑和结构关系，可以继续使用；案例内含的知识不属于计算机易于处理的内容，为便于健康管理一体化平台后续推理、检索等功能开发，需要进一步抽取知识。规则化方面，规则知识本身是规则化的结果，无需规则化描述；故障树可根据需求决定是否规则化；预案为非结构化数据需要规则化后使用，上述结论如表5-1所示。

表5-1 各类知识分析

|  |  |  |  |
| --- | --- | --- | --- |
| 现存主要知识类型 | 特点 | 是否进一步抽取知识 | 是否需要规则化 |
| 规则知识 | 因果关系突出，易于计算机处理 | 否 | 否 |
| 故障树 | 结构性强，上下级节点有严格逻辑关系，属于结构性知识 | 否 | 根据需求确定 |
| 预案 | 非结构化，属于过程性知识 | 是 | 是 |

5.2知识抽取方式分析

知识抽取是从结构化和非结构化的原始数据中抽取重构知识，它既是一种对于己有原始数据的重用，也是一种基于原始数据进行某种模式化知识的生成。从知识抽取过程的自动化程度来看，知识抽取可分为人工抽取、半自动抽取和自动抽取3种方式。不同抽取方式适用场景不同，下面将针对预案知识特性，推荐知识抽取方式。

原始记录的预案可能是较粗糙的语料数据，蕴含着大量的设备故障缺陷规律，还需要更进一步进行抽取以得到更细粒度的知识内容，人工智能子自然语言处理方面的研究应用取得了丰硕成果，因此推荐采用自动化抽取方式。

由于风洞领域知识专业性较强，应考虑大量的专业特点，形成一套具有针对性的文本分析流程。如何将领域知识进行合理组织，使其变成计算机可以理解的形式；如何基于该组织形式以及相关表达模式从文本中进行专业信息的抽取；如何对于抽取出的信息进行挖掘分析，都是该流程中关键的问题。

具体地，首先对于预案进行关键信息的抽取，初步将长文本信息抽象压缩为半结构化形式，一方面方便领域人员进行专业分析，同时也为后续的算法挖掘操作奠定基础。抽取的关键信息包括实体和关系。实体即各种事物、概念等对象，边表示节点之间的语义关系。抽取工具推荐双向长短时记忆网络（Bi-directional LSTM，BiLSTM）加条件随机场（Conditional Random Field，CRF）。

5.3规则化方法选型

知识规则化是贯穿知识库构建与应用全过程的关键问题。本节分析了故障诊断场景下故障树知识和预案知识的特点，作出了规则化方法的选用建议。

5.3.1故障树知识

故障树根据根据需求可分为直接应用和规则化后应用两种情形，直接应用面向操作人员根据故障树进行人工故障诊断的需求，而规则化后应用则面向故障自动推理功能开发的需求。下面以规则化后应用情形展开描述。

故障树知识规则化从以下两方面考虑：知识应用方面，基于故障树技术的故障诊断专家系统进行故障诊断时，推理机首先根据故障现象定位到故障树某一节点，然后遍历该节点所有的底事件，并根据各底事件的故障规则表达式判断该底事件是否发生故障。知识表示形式方面，一是故障树有较强的结构性，符合框架法形式。二是故障树的上级节点作为条件，下级节点作为结果，结果可能作为条件，条件可能存在复用，符合产生式形式。

综合上述考虑，采用框架和规则相结合的方法来进行知识的表示，利用框架来表示故障树结构，利用规则来表示故障树底事件的故障判断方法。

具体地，用故障树表来存储故障树结构及相关信息，故障树的每个节点对应一个框架，每个框架包括故障树编号、故障树名称、所属系统、父事件、故障判断规则、故障等级、故障现象和故障处置流程等属性。将每个框架的上述属性作为一条记录存储在知识数据库中形成框架表。用规则库表来存储故障树底事件的故障判断规则。规则库表中每条记录对应1条产生式故障判断规则，利用规则库表中的1条或者几条规则可以对故障树底事件的故障进行判断。

5.3.2预案知识

通过知识抽取技术从不同结构和类型的数据中提取出了计算机可理解和计算的结构化数据，以供进一步的分析和利用。通过从不同来源、不同结构的预案中获取知识，然后再对知识进行提取，得到结构化的知识并存入到知识图谱中。

经过知识抽取，获得的每个预案的属性应包括预案编号、预案名称、发生时间、发现人、故障原因、处置程序、预防措施等信息。进一步，需要对上述抽取到的信息进行融合。知识融合是将分散的知识组织形成一个高度互联的知识图谱，消除实体、关系、属性与事实对象之间的歧义，使不同来源的知识实现整合。针对同一属性出现不同值的情况，根据数据源的数量和可靠度进行判断，依据属性值的距离实现融合。下面推荐两种知识融合算法。

（1）基于余弦相似度的实体消歧算法

实体消歧算法，以试验系统设备知识库中两个同名实体作为输入，基于余弦相似度算法，输出两个同名实体的相似度，实现同名实体消歧功能，为试验系统设备知识库的统一化表达提供算法支持。余弦相似度通过直接计算TF矩阵中两个向量的相似度，实际上就是求解两个向量夹角的余弦值，即点乘积除以二者的模长：



这里的和分别代表*A*和*B*的各分量。将计算所得的相似度与设定阈值比较，低于阈值则代表两个同名实体指代内容不同，应当将两个实体都存入试验系统知识库中，否则应合并。

（2）基于Levenshtein距离的共指消解算法

共指消解算法，以试验系统设备知识库中两个非同名实体作为输入，基于 Levenshtein距离相似度算法，输出两个非同名实体的相似度，实现非同名实体的共指消解，为试验系统设备知识库的知识加工提供算法支持。

Levenshtein距离是一种编辑距离，指两个字串之间，由一个转成另一个所需的最少编辑操作次数。允许的编辑操作包括：将一个字符替换成另一个字符、插入一个字符和删除一个字符。利用求得的操作数计算字符串*A*和*B*之间的Levenshtein距离，计算该距离与两字符串中长度较小值之间的比值，作为基于Levenshtein距离的属性相似度，相似度达到设定的阈值，则代表两个非同名实体指代内容相同，可进行合并，从而达到共指消解目的。

5.4知识存储

知识存储环节可视为狭义的知识库，在整个知识库体系中起着承前启后的作用。知识存储模块的功能目标是对试验装备健康管理下的相关数据、知识进行梳理，并有效存储。本节根据故障树知识和预案知识的特点分别推荐了存储结构模型。

5.4.1故障树知识数据存储形式

为方便知识数据库的设计及管理，将故障知识数据用3个相关联的表表示，分别为故障监测表、故障树表和规则库表。故障知识数据3个知识表及综合数据库之间的相互关系如下：

（1）根据故障现象在故障监测表中获得故障节点编号；

（2）根据故障节点编号对故障树表进行检索，获得该故障节点的相关信息, 对该故障树遍历，获得该故障树所有底事件的故障判断规则索引；

（3）根据故障判断规则索引对规则库表进行检索，获得该底事件的故障判断规则表达式；

（4）根据故障判断规则表达式中的变量名称对综合数据库进行检索，获得该变量名对应的故障特征值。

故障树知识数据库E-R模型如图5-1所示。

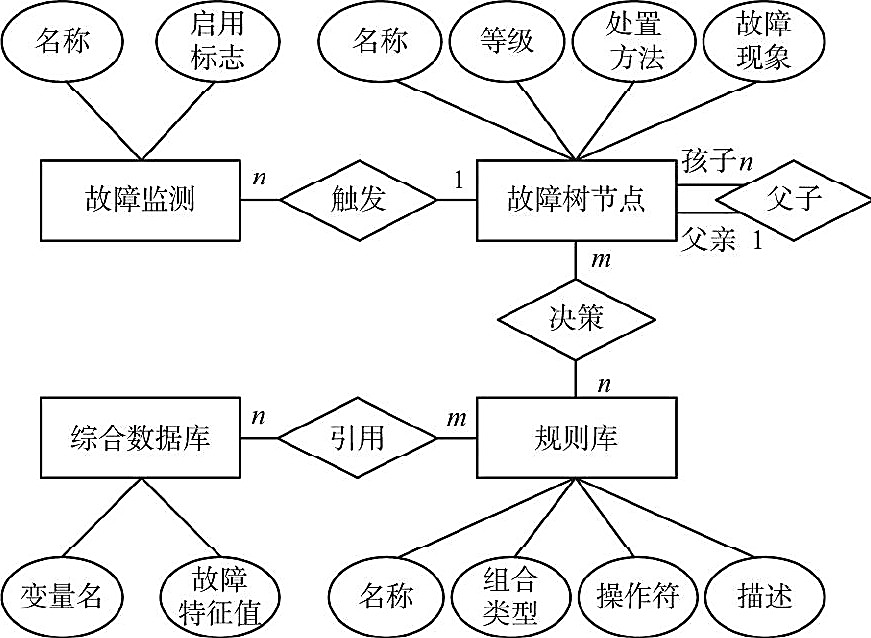


图5-1 故障树知识数据库E-R模型

5.4.2预案知识数据存储形式

图数据库直接使用图形结构来表示和存储数据，包含有节点，关系和属性（存储在节点或关系上的数据）。数据之间的关系视为与数据本身同等重要。图数据库中的每个节点（实体或属性）直接包含有与其他节点的关系的关系记录列表，访问时无需昂贵的搜索和匹配计算，因此图数据库擅长管理高度连接的数据和复杂查询。因此，推荐采用开源图数据库Neo4j作为图知识存储工具。

基于Neo4j的图形数据库模型如图3所示，网络化协作加工设备信息以及协作加工历史经验数据等以图形的形式存储，节点（node）与关系（relationship）构成有向图形，再将属性（property）赋予节点/关系，则构成了基于Neo4j的图形数据库模型。基于Neo4j的图数据库模型如图5-2所示。

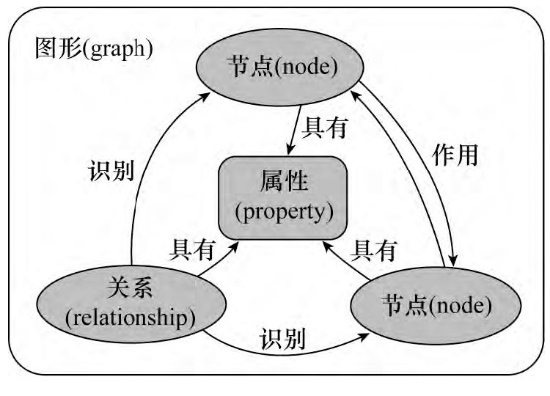


图5-2 基于Neo4j的图数据库模型

5.5知识维护

本章针对基地现行主要知识类型分析其维护内容并给出维护建议。规则知识需要考虑整体完备性，不完备的知识库将会导致推理紊乱。知识图谱的涵盖范围大，实体关系复杂多样，实体的增加与修改或实体间关系扩展必然造成整体知识图谱变化，需要频繁地对知识图谱进行更新与维护。预案知识维护的核心是，因此需要删除相似案例、老旧案例，减少冗余。

表5-2 知识维护内容表

|  |  |  |
| --- | --- | --- |
| 知识类型 | 维护需要 | 主要维护类型 |
| 规则知识 | 整体完备性 | 完备性维护、更新、备份 |
| 知识图谱 | 实体及实体关系更新 | 更新、备份 |
| 预案 | 以最小的资源消耗覆盖最多的故障类别 | 冗余性检测、更新、备份 |

5.5.1规则知识的维护

规则知识维护主要为完备性检验，库中的部分规则可能脱离整个规则网络而存在，其前提是不可触发的或结论是不可正视的，而使其永远不会被调用，对这样的规则，要设法修改使其进一步连通或者删除。

5.5.2知识图谱的维护

知识图谱的更新维护一般分为两个阶段，数据模式层的更新与数据层的更新。在数据模式层的更新过程中，可以对实体与实体间关系进行相关操作。通过数据模式层的更新补全算法能够做到实体的补齐，实体间关系的预测。在不改变知识图谱整体结构的基础上对实体属性进行增加修改操作。由于实体属性的删除容易破坏知识图谱的整体关系结构，所以在映射正确的前提下对相关属性进行删除操作。图数据库具备传统数据库所拥有的增、删、改功能，可实现上述维护。

5.5.3预案知识的维护

随着新问题的不断出现，存储于数据库中的新预案会越来越多，这就会造成一种预案库冗余现象，因此需要进行预案规模的限制。具体地，按照一定的规则对数据库中的预案进行分析比对，实现预案的增删改查，去除冗余的的预案，提高预案检索的精确性。

目前主要实现预案维护的方法是利用K-means聚类算法对预案库进行分析。它能在保持预案库覆盖率不降低的条件下，尽可能多地删除多余预案，以减少预案检索和预案修改的压力，同时考虑了预案推理系统的能力和效率，是预案库维护的一条有效途径。

参考文献

1. 刘鹏等. 知识表示与处理[M]. 北京：电子工业出版社, 2021:13-97
2. 朱小燕等. 人工智能知识图谱前沿技术[M]. 北京：电子工业出版社, 2020:17-54
3. Martins V W B, Rampasso I S, Anholon R, et al. Knowledge management in the context of sustainability: Literature review and opportunities for future research[J]. Journal of cleaner production, 2019, 229: 489-500.
4. 李向前. 复杂装备故障预测与健康管理关键技术研究[D]. 北京理工大学, 2014.
5. 陈文伟, 陈晟. 知识管理与知识工程[M]. 北京：清华大学出版社, 2016: 37-55
6. 姜达. 基于知识图谱的车载信号设备故障诊断研究[D]. 西南交通大学, 2020.
7. 张明宝, 施青青. 面向知识管理系统实现的软件基础设施研究[J]. 计算机集成制造系统, 2008(11): 2191-2198.
8. 宋晓霞. 变压器故障综合诊断专家系统的研究与实现[D]. 西安电子科技大学, 2010.
9. 闵鑫. 基于谓词逻辑的电网事故处理系统的设计与开发[D]. 华南理工大学, 2020.
10. 袁帅. 基于深度学习的知识库问答系统研究[D]. 电子科技大学, 2020.
11. DAY M-Y, TSAI R T-H, SUNG C-L, et al. Reference metadata extraction using a hierarchical knowledge representation framework[J]. Decision Support Systems, 2007, 43(1): 152–167.
12. Borkar V, Deshmukh K, Sarawagi S. Automatic segmentation of text into structured records[J]. ACM SIGMOD Record, 2001, 30(2).
13. 徐哲赢. 面向产品设计的跨平台知识集成技术研究[D]. 北京理工大学, 2016.
14. 李鹏翀. 基于领域语义的机构知识库检索系统的研发与应用[D]. 北京邮电大学, 2015.
15. 聂同攀, 曾继炎, 程玉杰, 马梁. 面向飞机电源系统故障诊断的知识图谱构建技术及应用[J]. 航空学报: 1-19.

附件1风洞设备知识管理案例

F1.1风洞设备故障诊断知识管理案例

根据《超高速试验设备故障汇编》，本案例运用产生式规则方法，对案例中的专家知识进行规则化描述，使得故障案例专家知识能够对风洞设备故障诊断任务的开展起到指导和协助的作用。

F1.1.1风洞本体

（1）故障案例记录（源知识）

①故障一：气动快速阀故障（DN250）

时间：2011年01月25日

故障发现人：李杰

故障/事故描述：阀门不能打开。

检查经过：现场检查。

发生原因：气动快速阀长期使用，阀座与轴套磨损严重，阀门前后压差较大时（>10MPa），快速阀不能打开。

解决措施：1.将快速阀前与总阀后的管道压力放空；2.拆卸阀门控制气源、气缸执行器、对阀门阀座进行维修更换；3.空载与负载试运行阀门开关位置、气密性。

经验教训：及时对阀门进行检修保养。

②故障二：闸板阀（Φ1.2m）脱落

时间：2011年03月10日

故障发现人：李杰

故障/事故描述：闸板阀阀板由于重力作用自行坠落。

检查经过：现场检查。

发生原因：阀门丝杠与阀板连接的铜螺母由于长期频繁开关使用，螺母磨损严重导致滑丝（螺母材质为铜），阀板由于重力作用自行坠落。

解决措施：1.将闸板阀前端试验段与后端真空球罐（5000m3）放空至大气状态；2.拆卸闸板阀上部电机，对闸板阀阀杆、阀板、轴承等拆卸清洗，更换密封圈、填料、丝杠螺母；3.对闸板阀进行手动盘车，调整上下限位后进行试车。

经验教训：为防止螺母磨损严重后，造成阀板脱落，将真空与高压管道非人为切断，酿成严重事故。根据试验运行次数，应定期拆卸闸板阀检查磨损情况并更换丝杠螺母。

③故障三：预充气快速阀（AQK1）开启异常

时间：2014年04月29日

故障发现人：吕超

故障/事故描述：预充气快速阀（AQK1）不能自动开启，将前后压差调整到10MPa能够正常打开。

检查经过：现场查看分析。

发生原因：拆卸发现阀门阀杆与阀盖连接垫磨损严重，阀座有明显磨损。

解决措施：1.将阀门前后压力放空，阀门整体拆卸后将执行器与阀体分离，对更换阀座、连接垫片、石墨填料；2.将阀门执行器与阀体安装后调试开关位置、检查气密性、空载与负载试运行。

经验教训：根据阀门使用工况，定期更换检修易损件。

④故障四：M4-8支路热阀故障

时间：2014年06月09日

故障发现人：吕超

故障/事故描述：热阀液压站在试验过时，油压升不到工作压力，液压站工作时异响。

检查经过：1.检查液压站油位、油质、蓄能器压力均正常；2.清洗液压油过滤网故障依然存在；3.拆卸液压站齿轮泵，发现齿轮泵内部密封垫已损坏；

发生原因：液压站齿轮泵损坏

解决措施：更换新齿轮泵，油站空载试运行热阀。

经验教训：1.定期更换易损坏橡胶密封圈2.定期紧固检查油路管道、齿轮泵、换向阀等关键部件状态。

⑤故障五：喷流预充气快速阀（PDC）密封不严

时间：2015年05月08日

故障发现人：吕超

故障/事故描述：喷流预充气快速阀（PDC）为电磁型快速开关阀门，在试验预充气过程中发现阀门关闭后，后段压力依然上升。

检查经过：1.检查发现阀门开关信号、位置均正常；2.对阀门前后管道放空，拆卸阀门发现内部有杂质导致阀门密封不严

发生原因：阀门内部有杂质

解决措施：对阀门各部件进行清洗后，空载与负载试运行阀门气密性。

经验教训：对使用率低的阀门，加大维护保养力度。

⑥故障六：DN250快速球阀关闭延迟故障

时间：2016年08月05日

故障发现人：吕超

故障/事故描述：快速阀在负载与空载关闭时出现滞后现象。

检查经过：检查信号、电磁阀、控制气源压力均正常，最终发现因阀门使用时间过长，导致关向快开阀磨损开启变慢，无法正常进行切换。

发生原因：阀门关向气动快开阀磨损，阀门使用时间过长，导致关向快开阀磨损开启变慢，无法正常进行切换。

解决措施：确定开关向快开阀接口及型号，对其进行更换并调整减压阀压力。

经验教训：对高负荷使用的阀门易损件定期进行检查保养。

（2）知识推理

由于《超高速试验设备故障汇编》已经对故障案例的记录内容进行了严格的规范，因此知识的录入模板仍然沿用该规范化模板。根据上述故障案例记录，人工对其中的知识进行抽取，运用产生式规则对知识中的规则进行规范化描述，所产生的规则化知识表示入下：

①IF GZ-FM001 and GZ-FM002 and GZ-FM003

THEN SS-FM001

②IF GZ-FM004 and GZ-FM005

THEN SS-FM002

③IF GZ-FM006 and GZ-FM007

THEN SS-FM003

④IF GZ-FM008 and GZ-FM009

THEN SS-FM004

⑤IF GZ-FM009

THEN SS-FM005

⑥IF GZ-FM010

THEN SS-FM006

其中，GZ-FM001表示条件“气动快速阀阀座磨损严重”，GZ-FM002表示条件“轴套磨损严重”，GZ-FM003表示条件“阀门前后压差较大（>10MPa）”，GZ-FM004表示条件“螺母磨损严重”，GZ-FM005表示条件“螺母滑丝”，GZ-FM006表示条件“阀杆与阀盖连接垫磨损严重”，GZ-FM007表示条件“阀座有明显磨损”，GZ-FM008表示条件“液压站齿轮泵损坏”，GZ-FM009表示条件“阀门内部有杂质”，GZ-FM010表示条件“阀门关向气动快开阀磨损”。SS-FM001表示结论“气动快速阀故障（DN250），阀门不能打开”，SS-FM002表示结论“闸板阀（Φ1.2m）脱落，闸板阀阀板由于重力作用自行坠落”，SS-FM003表示结论“预充气快速阀（AQK1）开启异常”，SS-FM004表示结论“M4-8支路热阀故障，油压升不到工作压力”，SS-FM005表示结论“喷流预充气快速阀（PDC）密封不严，试验预充气过程中发现阀门关闭后，后段压力依然上升”，SS-FM006表示结论“阀门关向气动快开阀磨损”。

F1.1.2运行控制系统

（1）故障案例记录（源知识）

①故障一：调温阀故障（M4-8支路）

时间：2015年08月28日

故障发现人：吕超

故障/事故描述：调温阀在试验过程中，命开度与实际开度不符

检查经过：试验后检查发现阀门卡住无法动作，执行器报警。1.对执行器的动力电、输入信号、反馈信号检测均正常；2.联系厂家对阀门现场故障码排除，仍无效。

发生原因：执行器过载

解决措施：更换换新电液缸执行器并进行参数调试。

经验教训：调温阀试用频率高、流量大、温度高、阀体易磨损，普通执行器无法满足使用工况，因此需更换功率较大的电液缸执行器。

②故障二：闸板阀（Φ1.2m）无法关闭

时间：2015年11月16日

故障发现人：吕超

故障/事故描述：闸板阀关行程时间过长，一直无法触发到位信号，控制柜热断路保护器跳闸，电机烧坏缺相。

检查经过：现场检查。

发生原因：下限为行程开关螺杆松动位移，电机持续不停动作，导致电机过载。

解决措施：1.对下限为行程开关的螺杆进行调整更换，增加弹簧垫片；2.更换备用电机并调试电机正反转；3.调整配电柜保护器保护电流与电机额定功率匹配；4.空载与负载试运行闸板阀工作状态。

经验教训：定期检查紧固闸板阀上下限位开关。

③故障三：试验过程总压一直升高

时间：2019年10月16日

故障发现人：崔炜栋

故障/事故描述：试验过程中，总压一直升高，直到停车。

检查经过：查看过程记录数据，总压反馈为0，现场查看，断线。

发生原因：总压传感器线断开。

解决措施：重新焊接断开点，上位软件增加断开自动检测功能。

经验教训：开车前注意检查相关信号状态、设备状况。

（2）知识推理

根据上述故障案例记录，人工对其中的知识进行抽取，运用产生式规则对知识中的规则进行规范化描述，所产生的规则化知识表示入下：

①IF GZ-FM001 THEN SS-FM001

②IF GZ-FM002 THEN SS-FM002

③IF GZ-FM003 THEN SS-FM003

其中，GZ-CS001表示条件“执行器过载”，GZ-CS002表示条件“下限为行程开关螺杆松动出现位移”，GZ-CS003表示条件“电机持续不停动作”，GZ-CS004表示条件“电机过载”，GZ-CS003表示条件“总压传感器线断开”。SS-CS001表示结论“调温阀故障（M4-8支路），命开度与实际开度不符”，SS-CS002表示结论“闸板阀（Φ1.2m）无法关闭或闸板阀关行程时间过长”，SS-CS003表示结论“试验过程总压持续升高，导致停车”。

F1.1.3加热器系统

（1）故障案例记录（源知识）

①故障一：加热器无绝缘

时间：2018年11月10日

故障发现人：陈久芬

故障/事故描述：试验结束后，试验当班人员发现加热器部分加热元件对地绝缘为零，无法送电。

检查经过：拆除热法，对加热器进行检查。检查发现：引电连接处绝缘陶瓷破碎，导致引电装置与内衬搭接短路，无法送电。

发生原因：1、加热器冷热交替工作，陶瓷强度有所下降。2、气流流量大，对上封头陶瓷冲击大，导致绝缘陶瓷破碎。

解决措施：由预热元件研制厂家对破碎的陶瓷件进行更换，加热器各组绝缘恢复正常，可正常送电，故障排除。

经验教训：应定期检查绝缘陶瓷，如有破碎及时更换。

②故障二：加热器无绝缘

时间：2019年8月10日，故障发现人：陈久芬

故障/事故描述：试验结束后，试验当班人员发现加热器部分加热元件对地绝缘为零，无法送电。

检查经过：拆除热法，对加热器进行检查。检查发现：引电导杆脱落后搭接到蓄热元件或者内衬上，导致加热器短路，无法送电。

发生原因：1、预热元件及引电导杆热膨胀量大，应力集中导致导电杆脱落。2、气流流量大，对导电杆冲击、震动大。

解决措施：由预热元件研制厂家对脱落的引电导杆进行更换或者重新焊接，加热器各组绝缘恢复正常，可正常送电，故障排除。

经验教训：1、加热器每次运行前检查绝缘电阻和内阻，发现异常开盖检修；2、对脱落导电杆重新焊接，确保焊接牢固。

（2）知识推理

根据上述故障案例记录，人工对其中的知识进行抽取，运用产生式规则对知识中的规则进行规范化描述，所产生的规则化知识表示入下：

①IF GZ-JR001 and GZ-JR002 and GZ-JR003 and GZ-JR004 and GZ-JR005

THEN SS-JR001

②IF GZ-JR006 and GZ-JR007 and GZ-JR008 and GZ-JR009 and GZ-JR010

THEN SS-JR001

综合①和②两条规则，可以产生如下规则：

IF [GZ-JR001 and GZ-JR002 and GZ-JR003 and GZ-JR004 and GZ-JR005] or [GZ-JR006 and GZ-JR007 and GZ-JR008 and GZ-JR009 and GZ-JR010]

THEN SS-JR001

其中，GZ-JR001表示条件“加热器冷热交替工作”，GZ-JR002表示条件“陶瓷强度下降”，GZ-JR003表示条件“气流流量大对上封头陶瓷冲击大”，GZ-JR004表示条件“引电连接处绝缘陶瓷破碎”，GZ-JR005表示条件“引电装置与内衬搭接短路”，GZ-JR006表示条件“预热元件热膨胀量大”，GZ-JR007表示条件“引电导杆热膨胀量大”，GZ-JR008表示条件“应力集中导致导电杆脱落”，GZ-JR009表示条件“气流流量大对导电杆冲击和震动大”，GZ-JR010表示条件“引电导杆脱落后搭接到蓄热元件或者内衬上”。SS-JR001表示结论“加热器无绝缘，加热器部分加热元件对地绝缘为零，无法送电”。

附件2案例编辑模板

为了便于从案例中抽取知识，在人工抽取知识前，首先应将故障案例按统一模板录入计算机。在故障案例中，故障的具体情况应尽可能详细的描述，以便获得完整的故障资料，包括故障现象、故障部位、解决方案等。下面给出故障案例的编辑模板。

①案例名：案例名由三部分组成，组合顺序与结构见图F2-1。

|  |  |  |  |
| --- | --- | --- | --- |
|  | ×× | ××× | ×××× |
| 故障发生所在单位 |  |  |  |
| 故障部位 |  |  |  |
| 故障描述 |  |  |  |

图F2-1 案例名结构图

②案例编号：案例编号由五部分组成，组合顺序与结构见图F2-2。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ×× | ××× | ××× | ×××× | ××× |
| 所属系统类别码 |  |  |  |  |  |
| 所属系统序号 |  |  |  |  |  |
| 装备类别码 |  |  |  |  |  |
| 案例发生日期 |  |  |  |  |  |
| 案例序号 |  |  |  |  |  |

图F2-2 案例编号结构图

③故障时间：精确到日期，格式参照：yyyy-mm-dd；

④故障发现人：略；

⑤故障原因：描述故障机理；

⑥故障部位：精确到故障系统的最小维修单元；

⑦故障现象：即现行规范“故障描述”条目；

⑧检查经过：按现行规范；

⑨处置方案：即现行规范的“解决措施”条目；

⑩预防和改进措施：即现行规范的“经验教训”条目。

案例编辑模板v1.0

|  |  |  |  |
| --- | --- | --- | --- |
| 故障案例名 |  | 案例编号 |  |
| 故障时间 |  | 故障发现人 |  |
| 故障部位 |  | | |
| 故障现象 |  | | |
| 检查经过 |  | | |
| 故障原因 |  | | |
| 处置方案 |  | | |
| 预防和改进措施 |  | | |

附件3知识管理案例

F3.1产生式规则的故障诊断专家系统应用案例

电力变压器故障诊断的方法很多，根据DL/T596-1996《电力设备预防试验规程》规定，常用的方法有绝缘试验、局部放电试验、绝缘油电气试验、油中溶解气体分析（Dissolved Gases Analysis，DGA）。其中，油中溶解气体分析法是诊断电力变压器内部故障最主要的技术手段。利用DGA数据来判断变压器故障的方法有很多种，如国际电工委员会（International Electrotechnical Commission，IEC）三比值法、Rogers比值法、特征气体法、气体比值图示法等。本案例基于IEC三比值法，运用产生式专家系统知识规则化方法，构建了变压器故障诊断专家系统。

（1）知识库设计

①变压器故障诊断三比值法

三比值法的原理是根据变压器内的油和绝缘物发生故障时产生气体成分含量的相对浓度与温度的相互依赖关系，从五种特征气体中选用溶解度扩散系数相近的气体成分组成三对比值，以不同的编码表示。表F3-1给出了三比值法的编码规则，表F3-2是对应的故障类别判断方法。三比值法根据表F3-1的编码规则和表F3-2的故障类型判断方法作为诊断故障性质的依据。

表F3-1 三比值法的编码规则

|  |  |  |  |
| --- | --- | --- | --- |
| 气体比值范围 | 比值范围编码 | | |
| C2H2/C2H4 | CH4/H2 | C2H4/C2H6 |
| <0.1 | 0 | 1 | 0 |
| [0.1，1） | 1 | 0 | 2 |
| [1，3） | 1 | 2 | 1 |
| ≥3 | 2 | 2 | 2 |

表F3-2 故障类型判断方法

|  |  |  |  |
| --- | --- | --- | --- |
| 编码组合 | | | 故障类型 |
| C2H2/C2H4 | CH4/H2 | C2H4/C2H6 |
| 0 | 0 | 1 | 低温过热（低于150℃） |
| 2 | 0 | 低温过热（150℃~300℃） |
| 2 | 1 | 中温过热（300℃~700℃） |
| 0，1，2 | 2 | 高温过热（300℃~700℃） |
| 1 | 0 | 局部放电 |
| 2 | 0，1 | 0，1，2 | 低能放电 |
| 2 | 0，1，2 | 低能放电兼过热热 |
| 1 | 0，1 | 0，1，2 | 电弧放电 |
| 2 | 0，1，2 | 电弧放电兼过热热 |

②知识的获取

本案例仅考虑变压器故障诊断知识的直接获取方式，通过和电力部门专家的交流，同时阅读分析相关专业文献，综合得到变压器故障的相关知识。通过系统设计的知识编辑器窗口对知识库中的知识进行输入。列举两条规则如下：

a.规则前提：气体比值C2H2/C2H4的编码为“0”，CH4/H2的编码为“1”，C2H4/C2H6的编码为“0”。规则结论：局部放电。

b.规则前提：气体比值C2H2/C2H4的编码为“2”，CH4/H2的编码为“0”，C2H4/C2H6的编码为“0”或“1”。规则结论：低能放电。

③知识表示方式

在选择知识表示方法的时候应该考虑：方法能否充分表示领域知识；是否有利于对知识进行推理；是否便于对知识的组织、维护和管理。一般的，具有经验性、因果性的知识适合用产生式表示法进行表示。产生式表示法有内在的优势。

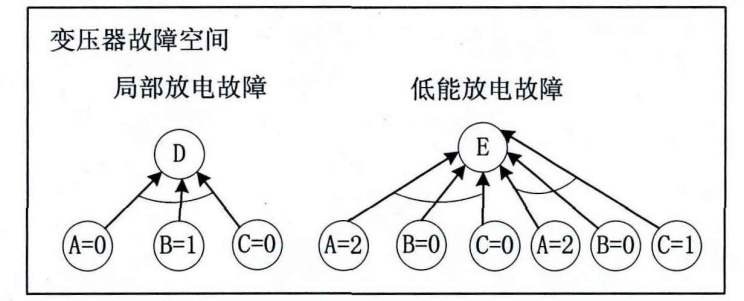
a.自然性：用“如果…，则…”的形式表示知识，与人类的判断性思维基本一致，既直观自然，又便于推理。

b.模块化：产生式规则是规则库中最基本的知识单元，它们同推理结构相对独立，且每条规则都具有相同的形式，有利进行模块化处理和知识的增、删、改，为规则库的建立和扩展提供了可管理性。

c.清晰性：产生式有固定的格式，每一条产生式规则都由前提和结论组成，所含的知识量都比较少，便于规则的设计和知识一致性和完整性的检测。

④“与或森林”知识表示方式

在三比值法中，要同时考虑三组气体比值C2H2/C2H4，CH4/H2，C2H4/C2H6，编码才能得出诊断结果。所以这三组气体比值对诊断结果的得出是“与”的关系，而每种诊断故障类型的得出对应的不只是一种编码的组合，这些编码组合对此种故障诊断得出又是“或”的关系。因此，用“与或森林”模型表示知识能够将故障知识的“与”、“或”关系表示的更为简单清晰。变压器故障特征到故障类型的推理之间没有中间状态，即变压器故障的知识空间是由高度为1的“与或树”组成。例如，在②中给出的两条知识规则对应的“与或森林”表示如图F3-1所示。



图F3-1 变压器知识空间“与或森林”

在图2-1中，为了简化标记，将比值C2H2/C2H4，CH4/H2，C2H4/C2H6的编码分为用A、B、C来表示，故障“局部放电”和“低能放电”分别用D和E表示。对应的产生式规则如下：

a.局部放电“与或树”对应的产生式规则为：

IF A为“0” and B为“1” and C为“0”

THEN 故障类型为D

b.低能放电“与或树”对应的产生式规则为：

·IF A为“2” and B为“0” and C为“0”

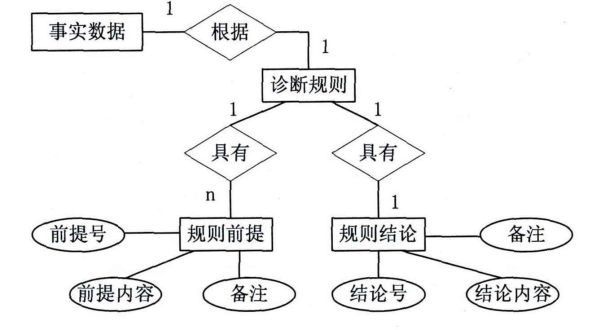
THEN 故障类型为E

·IF A为“2” and B为“0” and C为“1”

THEN 故障类型为E

⑤知识库概念结构设计

对于概念结构设计。选用易于表达和实现的实体——联系模型（Entity-Relationship，ER）。ER模型中用矩形来描述实体类型，用菱形来描述联系。本系统中涉及的实体有变压器故障的规则、规则前提、规则结论、事实数据。一个实体对应数据库逻辑设计中的一个表。实体之间的如图F3-2所示。



图F3-2 变压器知识库实体ER图

⑥知识库逻辑结构设计

将变压器知识库ER图转换成等价的关系模式后，故障知识库主要由规则表（cond\_conclusion），前提表（conditions），结论表（conclusion），事实表（facts）组成，如表F3-3、F3-4、F3-5和F3-6所示。规则表用于存储诊断规则，前提表和结论表分别存储规则的前提和结论，事实表用于存储用户提供的初始数据。

具有“与或”属性的规则都拆分作为独立的子规则存储，数据库中用type字段标识与或属性。type字段为0，表示此规则的条件是或属性；type字段为1，表示此规则是与属性。具体逻辑结构设计如下：

表F3-3 规则表（cond\_conclusion）逻辑结构

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名  Field\_name | 数据类型  type | 长度  length | 允许空  allow null | 备注  remark |
| id | int | 4 | no | 规则编号 |
| cond\_id | int | 4 | no | 规则前提编号 |
| conclusion\_id | int | 4 | no | 规则结论编号 |
| type | int | 4 | no | 规则关系编号 |

表F3-4 规则前提表（conditions）逻辑结构

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名  Field\_name | 数据类型  type | 长度  length | 允许空  allow null | 备注  remark |
| cond\_id | int | 4 | no | 规则前提编号 |
| condition | varchar | 100 | no | 规则前提内容 |
| remark | varchar | 100 | yes | 备注 |

表F3-5 规则结论表（conclusion）逻辑结构

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名  Field\_name | 数据类型  type | 长度  length | 允许空  allow null | 备注  remark |
| conclusion\_id | int | 4 | no | 规则结论编号 |
| conclusion | varchar | 100 | no | 规则结论内容 |
| remark | varchar | 200 | yes | 备注 |

表F3-6 事实表（facts）逻辑结构

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名  Field\_name | 数据类型  type | 长度  length | 允许空  allow null | 备注  remark |
| fact\_id | int | 4 | no | 事实编号 |
| C2H2 | float | 4 | no | C2H2含量 |
| C2H4 | float | 4 | no | C2H4含量 |
| C2H6 | float | 4 | no | C2H6含量 |
| CH4 | float | 4 | no | CH4含量 |
| H2 | float | 4 | no | H2含量 |
| cond\_id | int | 4 | no | 条件编号 |

（2）知识管理

知识管理的内容主要包括知识的添加、修改、删除和查询,以及对知识库中的知识进行一致性和完整性的检查等。

知识管理的主要功能如下：

a.知识的添加：提供对知识库中的知识不断扩充的途径。首先由知识工程师或用户从变压器专家和有关技术文献获取知识，经整理后由知识工程师通过知识编辑器输入到知识库中。

b.知识的查询：提供对变压器诊断规则、规则的条件和结论的查询功能，便于规则的定位和修改。

c.知识的维护：对知识库中的知识进行一致性和完整性的检验，所有的检查应在新规则加入时进行。

知识的一致性和完整性，主要表现在知识环路、冗余、从属等方面。在新的规则加入知识库时，需要检查新规则和规则库中已有规则是否形成环路、冗余和从属等。

a.环路规则：当一条规则链形成环路链时，如P→Q，Q→S，S→P，从任意一条规则进入推理，推理有可能陷入死循环。

检测方法：加入一条新规则时，在程序中用SQL语句检测。如规则库中已有规则P→Q，S>P，新加入规则为Q→S。检测步骤为：①在规则表F3-3中找出以S为条件的规则，记规则的结论为记录集L；②找出以Q为结论的规则，记规则的条件为记录集C；③若记录集L和C中有相同的元素，则存在环路，反之则表示无环路。

解决方式：当知识库中出现环路时，应征求专家的意见，修改或是舍弃其中的一条规则，破坏形成环路的条件。

b.冗余规则：指知识库中存在多余的知识或者存在多余的约束条件。

检测方法：加入一条新规则时，在程序中用SQL语句检测。如知识库中已有规则P→Q，Q>R，加入一条新规则P→R。检测步骤为:①在规则表F3-3中找出以P为条件的规则，记规则的结论为记录集L；②找出以R为结论的规则，记规则的条件为记录集C；③若记录集L和C中有相同的元素，则对规则库中已有的规则P→Q，Q→R而言，新规则P→R是冗余的，反之则新规则非冗余。

解决方式：若检测出新规则是冗余的，则不加入到知识库中。

c.从属规则：如果规则r1与r2，有相同的结论，但r1比r2要求更多的约束条件，则称r1是r2的从属规则。新规则加入时，需要征求专家意见处理规则。

（3）基于产生式规则的推理机设计

①推理控制策略简介

专家系统中的推理过程是一个专家的思维过程，求解问题的质量和效率不仅依赖于求解的方法，还依赖于求解的策略，这就是推理控制策略。推理控制策略主要包括推理方向、搜索策略和冲突消解策略。

②正向推理机设计

鉴于变压器三比值法故障诊断知识库中规则是精确的，采用基于产生式的正向推理效率较高。变压器基于产生式的正向推理是一种确定性推理方式，其基本思想是根据用户提供的己知事实出发，在知识库中选择出可用的知识，构成可用知识集,然后按照冲突消解策略从可用知识集中选出一条知识进行推理。如果规则的结论是另一条规则的条件，即条件和结论属于同一定义域，则把推出的结论当作新的事实存入数据库中，作为下一步推理的己知事实，循环进行推理直到得出所要求的解为止。

将现场获取的三组气体含量比值作为初始事实，在变压器故障知识库中选择匹配的知识后就能推理得出诊断故障类型。由于输出的是故障类型,输入的是气体比值，不是同一定义域，不需要考虑将得出的结论加入到事实表中。本系统是故障辅助诊断系统，期望给用户提供尽可能多的诊断供用户参考，所以不采用冲突消解策略，利用所有匹配的规则进行推理。具体的推理流程图如图2-3所示，推理步骤如下：

Step 1.从事实表facts（表F3-6）中提取数据，给初始参数C2H2，C2H4，CH4，H2，

C2H6赋值。

Step 2.计算气体比值x= C2H2/C2H4，y= CH4/H2，z= C2H4/C2H6，记为向量

R'=（x,y,z）。

Step 3.根据三比值法的编码规则表F3-1，将R'转化为对应的编码向量R=(X,Y,Z)，

其中X, Y, Z分别为气体比值x, y, z对应的编码。

Step 4.在表conditions（表F3-4）中搜索和R匹配的条件，记录条件编号为Cno

Cno = Search (R, Table condtions)

IF (Cno==NULL) // R is not in Table conditions

Goto Step 10

ELSE

Goto Step 5

Step 5.在表cond\_conclusion（表F3-3）中搜索和Cno匹配的所有规则，记为集合A，

A中规则的个数用K表示，第k条规则用A(k)表示

A = {(id, cond\_id, conclusion\_id, type) | cond\_id=Cno}

IF (K= 0)

Goto Step 10

ELSE

Goto Step 6

Step 6.处理匹配规则集A中的规则

FOR (k = 1,2,..K)

IF (A(k).type == 0)

Goto Step 7

ELSE

Goto Step 8

END FOR

Step 7.处理A中的“或”规则。将匹配规则集的结论记录到数组Result[i]，j=0

Result[j++]- A(k).conclusion\_id

Goto step 9

Step 8.处理A中的“与”规则。在表cond\_conclusion（表F3-3）搜索出和A集合中的第k条规则的结论和id 相同的规则，记为规则集B，B中规则的个数用N表示，第m条规则用B(m)表示。B = {(id, cond\_id, conclusion\_id, type) conclusion\_id= A(k). conclusion\_id &&id= A(k).id}

搜索规则集B中所有规则的条件是否在事实表facts（表F3-6）中

FOR(i=1,2,....N)

Index = Search(B(m).cond\_id, Table facts)

IF(Index == NULL)

Break

END IFEND FOR

IF(Index == NULL)

Goto Step 10

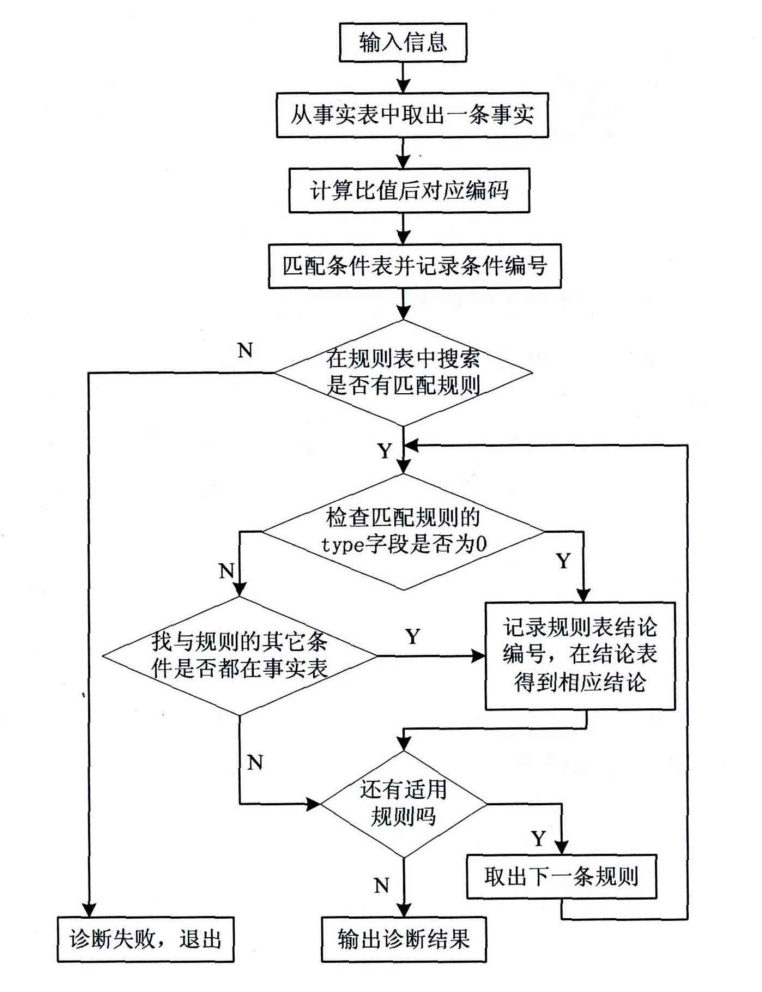
ELSE

Result[j++]= B(m). conclusion\_id

Step 9.在表conclusions(表F3-5)中搜索和Result[i]对应的结论，作为输出

{conclusionlconclusion\_id =Result[k], k—0, 1, ...., j}

Step 10.诊断失败。



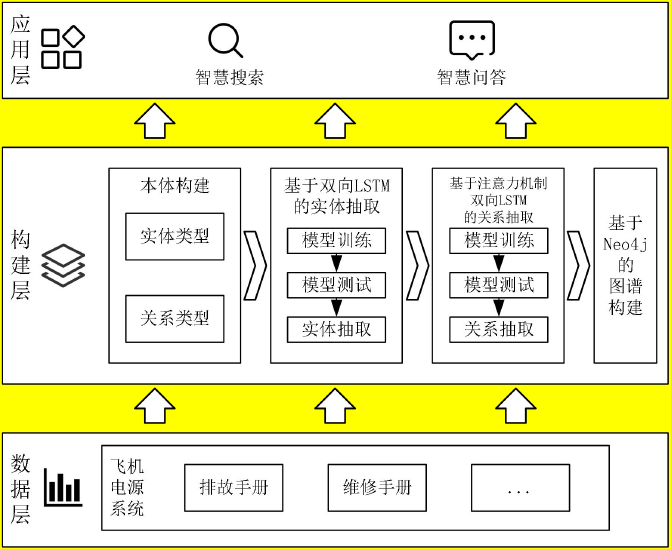
图F3-3 基于产生式规则的正向推理流程图

F3.2基于AI的故障诊断专家系统应用案例

面向飞机电源系统故障诊断方法虽然多种多样，然而常用的基于数据驱动的故障诊断方法无法利用专家知识，同时其结果可解释性差，为实际使用带来了不便。深度学习算法作为当前工智能领域的研究热点，具备将专家知识等非结构化数据进行结构化存储，并用于故障诊断的能力，能够实现对非结构化先验知识的利用以及故障原因的解释。

因此，本案例提出了一种面向飞机电源系统故障诊断的深度学习构建技术。首先，利用专家知识构建知识图谱的本体；然后，使用双向长短期记忆网络进行实体抽取；在此基础上，使用注意力机制双向长短期记忆网络进行关系抽取；最终利用抽取出的实体和关系构建面向飞机电源系统故障诊断的知识图谱。本案例以飞机电源系统故障排故手册为原始数据，对使用本文所提出的双向长短期记忆网络构建实体抽取方法进行了案例验证。

本案例提出的故障诊断方法包括本体构建、基于双向长短期记忆网络的实体抽取、基于注意力机制双向长短期记忆网络的关系抽取和基于Neo4j的知识图谱构建四个步骤，每一个步骤的具体内容如下：



图F3-4 面向飞机电源系统故障诊断方法的构建流程图

首先，根据故障诊断语料情况和图谱需求构建本体，确定实体类型和关系类型。

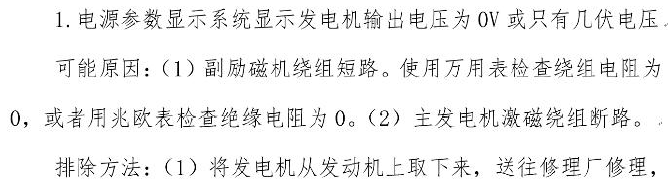
构建本体后，将故障诊断手册中部分语料分为训练集和测试集，对训练集和测试集进行实体标注。使用训练集训练基于双向LSTM的实体抽取模型，并使用测试集测试实体抽取模型效果。使用训练好的实体抽取模型抽取故障诊断手册语料中的实体。

对训练集和测试集进行关系标注。使用训练集训练基于注意力机制双向LSTM的关系抽取模型，并使用测试集测试关系抽取模型效果。结合实体抽取结果，使用训练好的关系抽取模型抽取故障诊断手册语料中的关系。

完成实体抽取和关系抽取后，使用知识图谱构建工具Neo4j，利用抽取出的故障诊断知识构建出面向故障诊断的知识图谱。

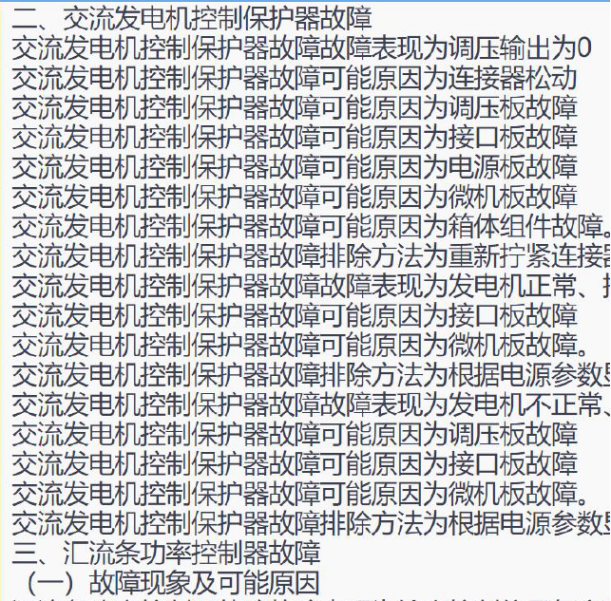
（1）数据介绍

本案例使用的数据来自飞机电源系统故障手册文档。该文档包含飞机电源系统可能发生的故障、故障表现、故障原因和解决措施。该文档部分内容如下图所示。



图F3-5 飞机电源系统故障手册文档部分内容

对飞机电源系统故障手册中的内容进行数据预处理。具体地，将“可能原因”“故障现象”“排除方法”之后的编号项中每一项与其对应的故障模式单独成句。经数据预处理后的飞机电源系统故障手册如下图所示。



图F3-6 处理后飞机电源系统故障手册文档部分内容

将该文档中部分语料划分为训练集和测试集，用于实体抽取和关系抽取模型的训练和测试。

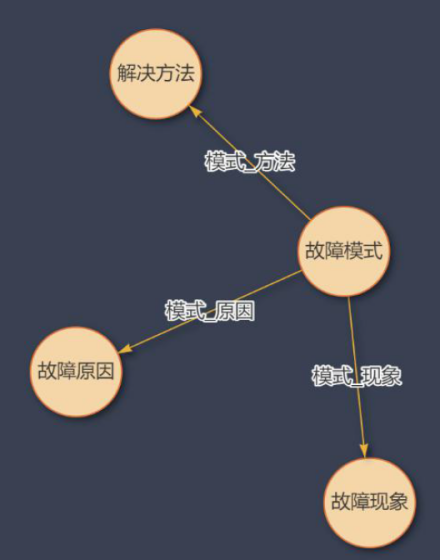
（2）本体构建

基于专家知识，本案例构建飞机电源系统故障诊断知识图谱的本体，共包含实体类型4个，关系类型3个。实体类型包含：故障模式、故障原因、故障现象和解决方法。关系类型及其头、尾实体见下表。

表F3-7 关系类型及其头尾实体

|  |  |  |
| --- | --- | --- |
| 关系类型 | 头实体类型 | 尾实体类型 |
| 模式\_原因 | 故障模式 | 故障原因 |
| 模式\_现象 | 故障模式 | 故障现象 |
| 模式\_方法 | 故障模式 | 故障方法 |

构建完成的本体可视化结果如下图所示。



图F3-7 本体可视化结果

（3）实体抽取结果

飞机电源系统故障手册文档原始语料划分出的训练集共有137句，4344字；测试集共有31句，910字。根据构建的本体标注出训练集中的实体，共标注出实体267个。实体标注结果如下图所示。



图F3-8 实体标注结果

例如，在上图中的原始语料中的句子“交流发电机控制保护器故故障表现为调压输出为0”中，标注出了故障模式实体“交流发电机控制保护器故障”和故障现象实体“调压输出为0”。将训练集文本组织成BMEO格式，处理后的训练集文本如下图所示。



图F3-9 处理后的实体抽取模型训练集文本

例如，在上图中，“交流发电机控制保护器故障”的实体类型为故障模式，因此其第一个字符“交”对应标签为“B-故障模式”，最后一个字符“障”对应标签为“E-故障模式”。

使用处理后的训练集文本训练双向LSTM实体抽取模型，设置参数批尺寸为32，时期为100。其中，所有训练样本在神经网络中进行一次正向传播和一次反向传播为一个时期。

进而，将测试集文本送入训练好的模型抽取实体，并与人工标注的实体进行对比。实体抽取模型测试结果如下图所示。



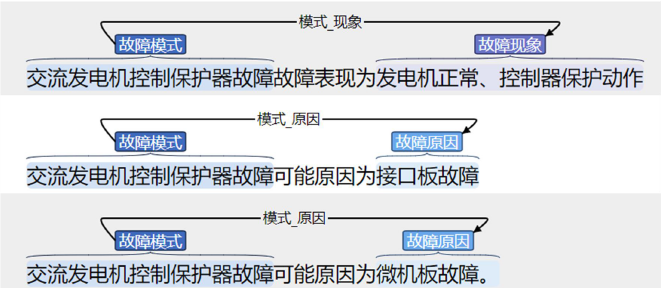
图F3-10 实体抽取模型测试结果

由上图可以看出，利用基于双向长短期记忆网络算法成功实现了对测试集文本语料的实体抽取，共910条知识，准确率为99.63%，召回率为71.58%。

完成测试后，使用实体抽取模型抽取原始文档中的实体，用于图谱构建。

（4）关系抽取结果

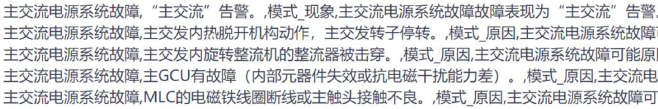
关系抽取训练集和测试集划分与实体抽取相同。根据构建的本体标注出训练集中的关系，关系标注结果如下图所示。



图F3-11 关系标注结果

例如，在上图中的原始语料中的句子“交流发电机控制保护器故障可能原因为接口板故障”中，标注出了实体“交流发电机控制保护器故障”和实体“接口板故障”之间的关系“模式\_原因”。

将测试集文本组织成“头实体尾实体关系所在句子”的格式，处理后的训练集文本如下图所示。



图F3-12 处理后的关系抽取模型训练集文本

例如，在上图中的句子“主交流电源系统故障表现为‘主交流’告警”中，实体“主交流电源系统故障”与实体“‘主交流’告警”之间的关系为“模式\_现象”，因此生成的语料尾“主交流电源系统故障主交流’告警模式\_现象主交流电源系统故障表现为‘主交流’告警”。

使用处理后的训练集文本训练基于注意力机制的双向LSTM关系抽取模型，设置参数嵌入层维度为100，位置向量字典大小为82，位置向量维度为25，隐藏层维度为200，批尺寸为1，时期为200，学习率为0.00005。

根据关系标注结果，将测试集文本组织成“头实体尾实体所在句子”的格式，将处理后的测试集文本送入训练好的模型抽取实体关系，并与人工标注的关系进行对比。关系抽取模型测试结果如下图所示。



图F3-13 关系抽取模型测试结果

由上图可以看出，利用基于注意力机制的双向长短期记忆网络算法成功实现了对测试集文本的关系抽取，共31条知识，准确率为75.00%，召回率为75.00%。完成测试后，使用原始文本的实体抽取结果将原始文本组织成“头实体尾实体所在句子”的格式，并使用关系抽取模型抽取原始文档中的关系，用于图谱构建。

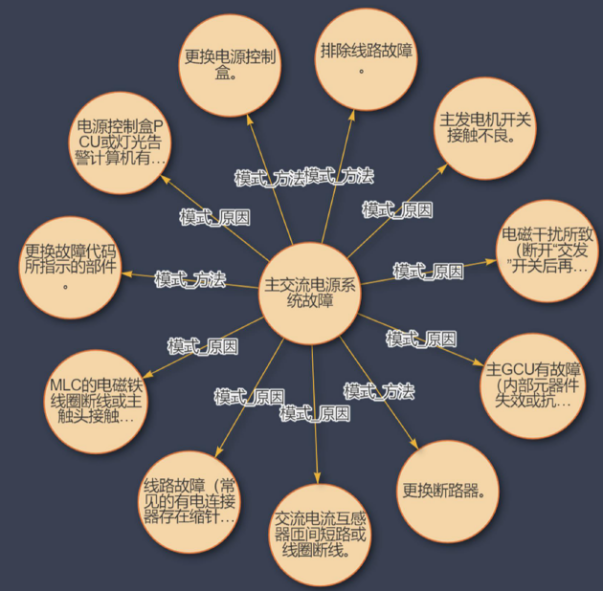
（5）知识图谱构建结果

原始语料中每个实体作为一个节点，每个关系作为一条连接其头实体和尾实体的边。使用Neo4j，将原始文本中人工标注和模型抽取的结果构建为飞机电源系统故障诊断知识图谱。所构建的知识图谱可视化结果如下图所示。



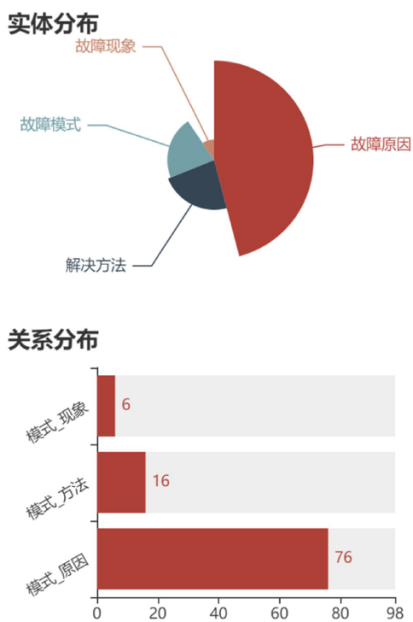
图F3-14 飞机电源系统故障诊断知识图谱可视化结果

构建的知识图谱因实体数较多，为清晰展示图谱中的实体和关系，截取该图谱部分实体与关系如下图所示。



图F3-15 部分飞机电源系统故障诊断知识图谱

该图谱包含实体74个，其中故障模式16个，故障原因34个，故障现象7个，解决方法17个。关系98条，其中模式\_原因76条，模式\_现象6条，模式\_方法16条。图谱详情如下图所示。



图F3-16 飞机电源系统故障诊断知识图谱详情

经人工审核，该图谱已将原始文档中大部分知识抽取并存储，实现了飞机电源系统故障诊断知识图谱的构建。随后，便可根据知识图谱形成大量的故障诊断规则，进一步开发相应的飞机电源系统故障诊断专家系统。